PATENT ABSTRACTS OF JAPAN

(11)Publication number : 11-025108
(43)Date of publication of application : 29.01.1999

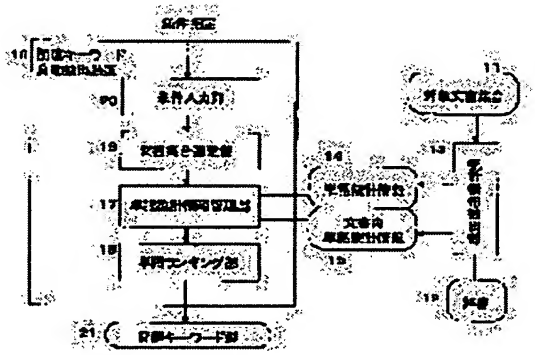
(51)Int.Cl. G06F 17/30

(21)Application number : 09-176822
(22)Date of filing : 02.07.1997

(71)Applicant : MATSUSHITA ELECTRIC IND CO LTD
(72)Inventor : SATO MITSUHIRO
NOGUCHI NAOHIKO
SUGANO YUJI
NOMOTO MASAKO
INABA MITSUAKI
FUKUSHIGE TAKAO

(54) AUTOMATIC EXTRACTION DEVICE FOR RELATIVE KEYWORD, DOCUMENT RETRIEVING DEVICE AND DOCUMENT RETRIEVING SYSTEM USING THESE DEVICES

(57)Abstract:
PROBLEM TO BE SOLVED: To automatically extract a relative keyword which is matched with the characteristics of a document to be practically retrieved and which is capable of obtaining one or more retrieval results at the time of executing retrieval using the keyword.
SOLUTION: An automatic extraction device for relative keywords is provided with a document set selection part 19 for specifying a partial set of each document based on the attribute information, input retrieval expression, etc., of the document, a word statistic information management part 17 for managing the statistic information of respective words in the whole objective document 11 and words appearing in each document as well as their statistic information 15; and a word ranking part 18 for calculating the importance of each word appearing in a partial set of a certain document and for aligning respective words in the order of importance, wherein the management part 17 quickly finds out the statistic information of respective words in the whole document and a specified partial set of the document. Consequently, words appearing in a certain document set can be ranked based on their importance and a part of the ranked words can be presented as a relative keyword.



LEGAL STATUS

[Date of request for examination] 22.02.2001
[Date of sending the examiner's decision of rejection]
[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]
[Date of final disposal for application]
[Patent number]
[Date of registration]

[Number of appeal against examiner's decision of rejection]

[Date of requesting appeal against examiner's decision of rejection]

[Date of extinction of right]

* NOTICES *

1.This document has been translated by computer. So the translation may not reflect the original precisely.
2.**** shows the word which can not be translated.
3.In the drawings, any words are not translated.

CLAIMS

[Claim(s)]
[Claim 1] Related keyword automatic-extracting equipment which is equipped with the following, extracts the group of a word or a word, and its significance, and is characterized by making it show in the form which can reuse this only about the particular part of the word groups which aligned. The document set selection section which specifies the subset of a document based on the reference formula which the attribute information by which statistical information, such as the frequency of occurrence of the group of the word which appears in each document of an object document set using a dictionary, or a word, and a distribution, was given to each document to the document set currently extracted beforehand, and the user inputted The word statistical information Management Department which manages the statistical information in the whole object document of each word and the word which appears in the document concerned for every document, and its statistical information The word ranking section which computes the significance of each word which appears in the subset specified based on the whole sentence document of each word, and the statistical information for every document, and aligns in order of significance
[Claim 2] The statistical information of the word which appears in the document group contained in Subset A when the subset B contained in this is specified by the document set selection section to the specified subset A in the aforementioned composition, Related keyword automatic-extracting equipment according to claim 1 characterized by computing the significance of each word which appears in Subset B, and being reflected in word ranking by seasoning the significance of each word in Subset B with difference with the statistical information of the word which appears in the document group contained in Subset B.
[Claim 3] Related keyword automatic-extracting equipment according to claim 1 or 2 characterized by computing the significance of the word concerned and being reflected in word ranking by seasoning the significance of the word which prepares the function which gives the weight of each document to the document set selection section, and is contained in each document of the specified document set with the weight of the document concerned.
[Claim 4] Related keyword automatic-extracting equipment according to claim 1 to 3 characterized by the ability to sort out only a word with high effectiveness in the case of reuse by excepting the word whose appearance degree is high frequency or low frequency in the whole object document set in consideration of the threshold which was able to be defined beforehand.
[Claim 5] Related keyword automatic-extracting equipment according to claim 4 characterized by the ability to sort out only a word with high effectiveness in the case of reuse by changing the threshold for exclusion according to characteristic quantity of the word, such as the length of a word.
[Claim 6] Related keyword automatic-extracting equipment according to claim 1 to 5 characterized by having the appearance Research and Data Processing Department which manages the information on the appearance position of a word, or the appearing context, computing the significance of the word concerned by considering the weight beforehand set to the significance of a word according to the kind of appearance information on the word, and being reflected in word ranking.
[Claim 7] Related keyword automatic-extracting equipment according to claim 1 to 6 characterized by for the part of speech of a word etc. having the language attribute Management Department which manages the attribute information on each word, computing the significance of the word concerned by considering the weight beforehand defined according to the attribute of the word concerned, and being reflected in word ranking.
[Claim 8] The inclusion relation as a character string between the extracted words or the word group specified

beforehand, and the extracted word When judged with having the character string inclusion relation judging section judged according to the defined conditions, and the words concerned having the inclusion relation as a character string The specified conditions are followed. only the character string of a long unit only the character string of a short unit Only a character string with a higher significance or by choosing whether they are the both sides of the difference of the character string of a short unit and the character string of a long unit, and the character string of a short unit, and ******** Related keyword automatic-extracting equipment according to claim 1 to 7 characterized by the ability to sort out only a word with high effectiveness in the case of reuse.

[Claim 9] Related keyword automatic-extracting equipment according to claim 1 to 8 characterized by the ability to classify and show the word extracted by the part of speech of a word etc. having the language attribute Management Department which manages the attribute information on each word, and taking into consideration the frequency of occurrence in the attribute of the word concerned, and the specified whole subset or a whole document, a distribution, etc.

[Claim 10] Related keyword automatic-extracting equipment according to claim 9 with which only the representation word group which prepares the representation word grant section which gives the word representing the set about each of the classified word group, and represents the classified word group is characterized by the ability to show a representation word and all words.

[Claim 11] As opposed to the document set from which statistical information, such as the frequency of occurrence of the group of the word which appears in each document of an object document set using a dictionary, or a word, and a distribution, is extracted beforehand The reference condition input section which inputs conditional expression required for a document retrieval, and the document-retrieval section which searches a document from an object document set according to the inputted reference conditions, About the document searched in the document-retrieval section 45, have the document ranking section 46 which calculates the goodness of fit between the reference formulas and documents which were inputted, and it changes. Document-retrieval equipment which can input into the reference condition input section the related keyword which sent the ranking result in the document ranking section to related keyword automatic-extracting equipment, and was fed back from related keyword automatic-extracting equipment.

[Claim 12] It is document-retrieval equipment possible in having the document-retrieval section which searches a document from an object document set according to the reference conditions inputted as the reference condition input section which inputs conditional expression required for a document retrieval, changing, and the aforementioned reference condition input section inputting considering the related keyword in which the user has been seen off from related keyword automatic-extracting equipment in addition to inputting reference conditions as reference conditions.

[Claim 13] As opposed to the document set from which statistical information, such as the frequency of occurrence of the group of the word which appears in each document of an object document set using a dictionary, or a word, and a distribution, is extracted beforehand The reference condition input section which inputs conditional expression required for a document retrieval, and the document-retrieval section which searches a document from an object document set according to the inputted reference conditions, The document-retrieval equipment which has the document ranking section 46 which calculates the goodness of fit between the reference formulas and documents which were inputted, and changes about the document searched in the document-retrieval section 45, It consists of related keyword automatic-extracting equipment connected to the aforementioned document-retrieval equipment. The ranking result outputted from the document ranking section of the aforementioned document-retrieval equipment is sent to related keyword automatic-extracting equipment. Moreover, the document-retrieval system characterized by feeding back a related keyword to the reference condition input section of related keyword automatic-extracting equipment to document-retrieval equipment, and performing retrieval by keyword.

[Claim 14] The ranking result which the document set selection section was prepared between document-retrieval equipment and related keyword automatic-extracting equipment, and was outputted from the document ranking section of document-retrieval equipment is a document-retrieval system according to claim 13 characterized by to be sent to the document set selection section, to perform specification of a document, and to input into the aforementioned related keyword automatic-extracting equipment 48 the subset of the document which the document set selection section 47 specified.

[Claim 15] The document-retrieval system according to claim 13 or 14 characterized by using related keyword automatic-extracting equipment according to claim 1 to 10 for related keyword automatic-extracting

equipment.
[Claim 16] The document-retrieval equipment which has the document-retrieval section which searches a document from an object document set according to the reference conditions inputted as the reference condition input section which inputs conditional expression required for a document retrieval, and changes, It consists of related keyword automatic-extracting equipment connected to the aforementioned document-retrieval equipment. the reference condition input section of the aforementioned document-retrieval equipment The document-retrieval system characterized by inputting the related keyword which has been sent from related keyword automatic-extracting equipment in addition to a user inputting reference conditions as reference conditions, and performing retrieval by keyword.
[Claim 17] The document-retrieval system according to claim 16 characterized by using related keyword automatic-extracting equipment according to claim 1 to 10 for related keyword automatic-extracting equipment.

[Translation done.]

* NOTICES *

1.This document has been translated by computer. So the translation may not reflect the original precisely.
2.**** shows the word which can not be translated.
3.In the drawings, any words are not translated.

DETAILED DESCRIPTION

[Detailed Description of the Invention]
[0001]
[Industrial Application] this invention relates to the related keyword automatic-extracting equipment for extracting as a keyword the phrase by which the document set is characterized from a specific document set, and the document-retrieval equipment using the aforementioned related keyword automatic-extracting equipment.
[0002]
[Description of the Prior Art] In document-retrieval equipment, although it is necessary to input the reference formula using the suitable search term in order to obtain the document which a user needs, there is a problem that the user itself cannot recollect a suitable search term easily. Then, the technique of helping re-reference of a user etc. has been taken by showing the word relevant to a search term conventionally to the search term which the user inputted using a related-term dictionary etc. However, in order to depend for such technique on the property of a ** useless **** related-term dictionary statically beforehand, the related term adapted to the property of the document used as the candidate for reference is not obtained. Moreover, there was a fault that it was not guaranteed that at least one or more documents are obtained as a result of referring to the obtained word.
[0003]
[Problem(s) to be Solved by the Invention] Statistical information, such as the frequency of occurrence, a distribution, etc. of each word in the document set which this invention solves the aforementioned technical problem and was specified, By computing the significance of a word in consideration of the statistical information of the word in the whole document for reference, carrying out ranking of the word with the significance based on this, and extracting the word group which is a part of rank It aims at offering the related keyword automatic-extracting equipment which is based on the property of the actual document for reference, and can extract the high related keyword group of quality at high speed and dynamically.
[0004] Moreover, when reference is performed using the related keyword group obtained from the aforementioned related keyword automatic-extracting equipment, it aims at offering the document-retrieval system using the document-retrieval equipment and these which guarantee that at least one or more reference results are obtained.
[0005]
[Means for Solving the Problem] this invention in order to attain the above-mentioned purpose as related keyword automatic-extracting equipment The document set selection section which specifies the subset of a document based on the reference formula which the attribute information given to each document and the user inputted, With the word which appears for every statistical information in the whole object document of each word, or document, and the word statistical information Management Department which manages the statistical information The word ranking section which computes the significance of each word which appears in the subset of the document specified based on the whole sentence document of each word or each statistical information in a document, and aligns in order of significance is prepared. by the word statistical information Management Department It is possible to ask for the whole document and the statistical information of each word in the specified document subset at high speed, ranking of each word which appears in the specified document set can be carried out at high speed in order of the significance, and the part can be shown as a related keyword.
[0006] In the aforementioned composition, furthermore, by in addition, the thing for which the word which the

weight of a word is changed or fulfills specific conditions from the word group after ranking by establishing a means to manage the appearance position in the attribute information on a word or a document etc. is deleted More intelligible related keyword presentation can be performed by classifying the word group which the precision as a related term of the word group extracted could be raised, and was extracted according to the attribute and statistical property of a word.
[0007] Moreover, by this invention's constituting the document-retrieval system which contains the document-retrieval equipment which cooperated with related keyword automatic-extracting equipment in order to attain the above-mentioned purpose, and reusing the extracted related keyword as an input The extracted related keyword suits the property of an object document, and if the candidate for reference is the same document group, since it is secured that at least one or more reference results are obtained by the keyword, re-reference can be performed efficiently and easily.
[0008]
[Embodiments of the Invention] As opposed to the document set from which statistical information, such as the frequency of occurrence of the group of the word to which invention of this invention according to claim 1 appears in each document of an object document set using a dictionary, or a word, and a distribution, is extracted beforehand The document set selection section which specifies the subset of a document based on the reference formula which the attribute information given to each document and the user inputted, With the statistical information in the whole object document of each word and the word which appears in the document concerned for every document, and the word statistical information Management Department which manages the statistical information It has the word ranking section which computes the significance of each word which appears in the subset specified based on the whole sentence document of each word, and the statistical information for every document, and aligns in order of significance. Only about the particular part of the word groups which aligned, the group of a word or a word, and its significance is extracted, and it has operation of carrying out high-speed presentation in the form which can reuse this.
[0009] Invention of this invention according to claim 2 is set to related keyword automatic-extracting equipment according to claim 1. The statistical information of the word which appears in the document group contained in Subset A when the subset B contained in this is specified by the document set selection section to the specified subset A, The significance of each word which appears in Subset B is computed, and it is made to be reflected in word ranking by seasoning the significance of each word in Subset B with difference with the statistical information of the word which appears in the document group contained in Subset B.
[0010] Invention of this invention according to claim 3 computes the significance of the word concerned, and it is made to reflect it in word ranking in related keyword automatic-extracting equipment according to claim 1 or 2 by seasoning the significance of the word which prepares the function which gives the weight of each document to the document set selection section, and is contained in each document of the specified document set with the weight of the document concerned.
[0011] Invention of this invention according to claim 4 enables it to sort out only a word with high effectiveness in related keyword automatic-extracting equipment according to claim 1 to 3 in the case of reuse by excepting the word whose appearance degree is high frequency or low frequency from the object of related keyword extraction in the whole object document set in consideration of the threshold which was able to be defined beforehand.
[0012] Invention of this invention according to claim 5 enables it to sort out only a word with high effectiveness in related keyword automatic-extracting equipment according to claim 4 in the case of reuse by changing the threshold for exclusion according to characteristic quantity of the word, such as the length of a word.
[0013] Invention of this invention according to claim 6 has the appearance Research and Data Processing Department which manages the information on the appearance position of a word, or the appearing context, and it computes the significance of the word concerned and it is made to reflect it in word ranking in related keyword automatic-extracting equipment according to claim 1 to 5 by considering the weight beforehand set to the significance of a word according to the kind of appearance information on the word.
[0014] The part of speech of a word etc. has the language attribute Management Department which manages the attribute information on each word, and invention of this invention according to claim 7 computes the significance of the word concerned, and it is made to reflect it in word ranking by considering the weight beforehand defined according to the attribute of the word concerned in related keyword automatic-extracting equipment according to claim 1 to 6.
[0015] Invention of this invention according to claim 8 is set to related keyword automatic-extracting

equipment according to claim 1 to 7. The inclusion relation as a character string between the extracted words or the word group specified beforehand, and the extracted word When judged with having the character string inclusion relation judging section judged according to the defined conditions, and the words concerned having the inclusion relation as a character string The specified conditions are followed. only the character string of a long unit only the character string of a short unit Or only a character string with a higher significance enables it to sort out only a word with high effectiveness in the case of reuse by choosing whether they are the both sides of the difference of the character string of a short unit and the character string of a long unit, and the character string of a short unit, and ********.

[0016] Invention of this invention according to claim 9 classifies the extracted word, and enables it to show it in related keyword automatic-extracting equipment according to claim 1 to 8 by the part of speech of a word etc. having the language attribute Management Department which manages the attribute information on each word, and taking into consideration the frequency of occurrence in the attribute of the word concerned, and the specified whole subset or a whole document, a distribution, etc.

[0017] Only the representation word group which prepares the representation word grant section which gives the word representing the set about each of the classified word group, and represents the classified word group enables it, as for invention of this invention according to claim 10, to show a representation word and all words in related keyword automatic-extracting equipment according to claim 9.

[0018] As opposed to the document set from which statistical information, such as the frequency of occurrence of the group of the word to which invention of this invention according to claim 11 appears in each document of an object document set, using a dictionary as document-retrieval equipment, or a word, and a distribution, is extracted beforehand The reference condition input section which inputs conditional expression required for a document retrieval, and the document-retrieval section which searches a document from an object document set according to the inputted reference conditions, It has the document ranking section which calculates the goodness of fit between the reference formulas and documents which were inputted about the document searched in the document-retrieval section. It has operation of inputting into the reference condition input section the related keyword which sent the ranking result in the document ranking section to related keyword automatic-extracting equipment, and was fed back from related keyword automatic-extracting equipment.

[0019] The reference condition input section which inputs the conditional expression [a document retrieval] as document-retrieval equipment to be invented [of this invention / according to claim 12], It has the document-retrieval section which searches a document from an object document set according to the inputted reference conditions. the aforementioned reference condition input section It has operation of inputting the related keyword which has been sent from related keyword automatic-extracting equipment in addition to a user inputting reference conditions as reference conditions.

[0020] As opposed to the document set from which statistical information, such as the frequency of occurrence of the group of the word to which invention of this invention according to claim 13 appears in each document of an object document set, using a dictionary as a document-retrieval system, or a word, and a distribution, is extracted beforehand The reference condition input section which inputs conditional expression required for a document retrieval, and the document-retrieval section which searches a document from an object document set according to the inputted reference conditions, The document-retrieval equipment which has the document ranking section which calculates the goodness of fit between the reference formulas and documents which were inputted, and changes about the document searched in the document-retrieval section, It has related keyword automatic-extracting equipment connected to the aforementioned document-retrieval equipment. It has operation of sending the ranking result outputted from the document ranking section of the aforementioned document-retrieval equipment to related keyword automatic-extracting equipment, and feeding back a related keyword to the reference condition input section of related keyword automatic-extracting equipment to document-retrieval equipment, and performing retrieval by keyword.

[0021] Invention of this invention according to claim 14 is set to a document-retrieval system according to claim 13. The document set selection section is prepared between document-retrieval equipment and related keyword automatic-extracting equipment. The ranking result outputted from the document ranking section of document-retrieval equipment is sent to the document set selection section, specification of a document is performed, and the subset of the document which the document set selection section 47 specified is inputted into the aforementioned related keyword automatic-extracting equipment 48.

[0022] In a document-retrieval system according to claim 13 or 14, as for invention of this invention according

to claim 15, related keyword automatic-extracting equipment according to claim 1 to 10 is used for related keyword automatic-extracting equipment.
[0023] The reference condition input section which inputs conditional expression [a document retrieval] to be invented [of this invention / according to claim 16], The document-retrieval equipment which has the document-retrieval section which searches a document and consists of an object document set according to the inputted reference conditions, It has related keyword automatic-extracting equipment connected to the aforementioned document-retrieval equipment. the reference condition input section of the aforementioned document-retrieval equipment It has operation of inputting the related keyword which has been sent from related keyword automatic-extracting equipment in addition to a user inputting reference conditions as reference conditions, and performing retrieval by keyword.
[0024] In a document-retrieval system according to claim 16, as for invention of this invention according to claim 17, related keyword automatic-extracting equipment according to claim 1 to 10 is used for related keyword automatic-extracting equipment.
[0025] Below, the form of concrete operation of this invention is explained with reference to an attached drawing.
[0026] (Form 1 of operation) The form of operation of the 1st of this invention is explained to the beginning. Drawing 1 is the block diagram having shown the composition of the related keyword automatic-extracting equipment concerning the form of operation of the 1st of this invention. First, the statistical information extraction section 13 which operates as pretreatment extracts the word statistical information 14, such as a frequency distribution of the word in the whole document set, and the word statistical information 15 in a document which is the statistical information of the word contained in the document concerned for every document to the target document set 11 using the dictionary 12. Drawing 2 (a) is the table format view showing the structure of word statistical information, and drawing 2 (b) is the table format view showing the structure of the word statistical information in a document. The word statistical information 14 is stored as a table as shows the statistical information of the word extracted by the statistical information extraction section 13 to drawing 2 (a). By using this table, it can ask for the full force present frequency and the number of appearance documents of a word "the Internet" whole sentence in the letter at high speed. Moreover, the word statistical information 15 in a document is stored as a table as shows the statistical information of the word for every document to drawing 2 (b). Thereby, a word "the Internet" can ask a publication number 0010 for the statistical information for every document that a word "WWW" appears twice, at high speed 5 times.
[0027] Related keyword automatic-extracting equipment 16 consists of the word statistical information Management Department 17 which manages the word statistical information 14 of the whole document, and the word statistical information 15 in a document, the word ranking section 18 which computes the significance of a word, the document set selection section 19 which specifies the subset of an object document, and the condition input section 20 which inputs the selection conditions to the document set selection section 19 and which is a means.
[0028] Operation of related keyword automatic-extracting equipment 16 which has this composition is explained below. First, the document set selection section 19 specifies a document set according to the conditions inputted to the condition input section 20. A document set is specified with either of three kinds of meanses as follows, or its combination.
(1) Specify a document set according to the attribute of a document. In this case, the genre to which a document belongs has a means to choose a document by the attribute value beforehand given to the document, and the document set selection section 19 adopts as a subset the document group corresponding to the attribute value specified by the condition input section 20.
(2) Specify a document set by the reference formula. In this case, the document set selection section 19 has a document-retrieval means to specify the document which suits the reference formula inputted in the condition input section 20, and adopts as a subset the document group obtained using this as a result of reference. In addition, if there is a function which judges a goodness of fit with a reference formula for a document-retrieval means, and carries out ranking of the document to the order of a goodness of fit in that case, you may adopt the particular part of the reference results, for example, high order 10 document, as a subset.
(3) The document set specified by the user. In this case, the document set selection section 19 adopts as a subset the document (plurality) which the user specified directly in the condition input section 20.
[0029] The document set selection section 19 passes the word statistical information Management Department 17 the document set selected by the above as the set of publication numbers of the identifier which

determines each document as a meaning, for example, a list. To the specified document set, the word statistical information Management Department 17 investigates the word statistical information 14 in a document from a publication number for every document, and gets the frequency of occurrence in the word which appears in the document concerned, and each document. Next, the word statistical information 15 is investigated about all the obtained words, and the frequency and distribution information in a whole sentence document on the word concerned are acquired.
[0030] The various statistical information obtained here is passed to the word ranking section 18, and the significance of each word is computed. Significance [of a certain word "W"] S (W) is computable as follows, for example.
[Equation 1]

$$S(W) = C * \sum_{j=0}^{n} \{TF_j(W) * IDF(W)\} * FN(W)$$

However, C : Constant n : The number TFj of documents contained in the specified document set (W): Document Dj The frequency of occurrence FN of the word "W" which can be set (W) : It is the number of documents which are document [which was specified] gathering and contains the word "W."
[0031] Moreover, IDF (W) is an index called idf value of the word "W", for example, is calculated by the following formulas.
IDF(W)= 1−log(DF(W)/N)
However, DF (W): The number N of documents in which the word "W" appears in the whole document : It is the number of whole sentence documents.
[0032] As for IDF (W), the value becomes small when the word "W" appears in more documents (that is, it is a more general word). Thereby, the significance of the word which appears comparatively well in the whole object document can be suppressed low. the significance of the word which appears in the specified document set mostly by furthermore taking FN (W) into consideration -- high -- it can do -- a result -- a significance high into a word characteristic of the specific document set -- it can give . In addition, in the above−mentioned computing method, you may normalize TF (W) by document sizes (the character number, the number of differences of a word contained) of a document, full force present frequency of a word, etc. in which the word is contained.
[0033] The word ranking section 18 performs significance calculation about all the words contained in the whole sentence document in the specified subset, and aligns all words in order of significance after that. At the end, a particular part, for example, high order 10 word, is adopted from the word group which aligned, and it shows as a group of a word or a word, and its significance. In addition, you may show simultaneously the various statistical information used not only for significance but for significance calculation on the occasion of extraction. Moreover, the group of the extracted related keyword and its significance can also be accumulated as a user's history. By doing in this way, the large application of becoming possible to express a range, taste, etc. of interest of a user as a vector of a keyword and its weight, and using this vector for other operations, for example, reference of a document set, etc. is possible.
[0034] If the above formula is used, it can carry out like the example shown, for example in drawing 3 , and related keyword automatic extracting can be performed. This drawing 3 is drawing showing the flow of the procedure of related keyword automatic−extracting operation. In drawing 3 , the word statistical information Management Department 17 by which the publication−number list 31 was inputted outputs the word which appears in the corresponding publication number (for example, 0010, 0341 grades), and its frequency for every document, and gets the word statistical information 33, 34, and 35 in a document. Simultaneously, the statistical information 32 whole sentence in the letter is obtained to all the words called for here. Next, such statistical information 32, 33, 34, and 35 is passed to the word ranking section 18. In the word ranking section 18, the significance of each word is calculated based on the various statistical information 32−35 using the aforementioned formula. It is as follows when it is the case of drawing 3 (however, C is set to 1 and N is set to 10000).
IDF(applet) = 1−log (86/10000)
= 5.756 S (applet) = 2*5.756+6*5.756*2 = 92.096 IDF(Internet) = 1−log (1129/10000)
= 3.181 S(Internet) = (3*3.181+1*3.181+2*3.181) *3 = 57.258 IDF (CGI) = 1−log (79/10000)

= 5.840 S(CGI) =(4*5.756)*1 = 23.024 IDF(WWW) = 1-log(615/10000)
= 3.789 S(WWW) =(5*3.789)*1 = 18.945 IDF(JAVA) = 1-log(161/10000)
= 5.129 6 S(JAVA) =(6*5.129+3*5.129+3*5.
129)*3 = 184.644 IDF(SUN) = 1-log(35/10000)
= 6.655 S (SUN) =(6*6.655) *1 = 39.930 IDF(script) = 1-log (813/10000)
= 3.510 S (script) =(5*3.510) *1 = 17.550 [0035] In the word ranking section 18, a word is aligned with the significance searched for as mentioned above, and the word list 37 of [after alignment] is obtained. Here, if it has been specification that three high orders of the word by which ranking was carried out are extracted, "JAVA" which is three high orders in the word list 37, an "applet", and the "Internet" are extracted as a related keyword.
[0036] Although it came above as an object of extraction of one word registered into the dictionary, generally the group of not only a word but a word is sufficient. The group of a word points out the group of the compound constituted by continuation of a noun, and the noun connected with particle "", the group of the noun connected with a particle "**" and "**", and a verb, etc. If such statistical information can be extracting in advance like a word, the technique shown above can apply as it is, and the group of a word can be extracted as a related keyword.
[0037] In addition, the related keyword input unit 16 is good also considering the document set selection section 19 and the condition input section 20 as another composition. When the document set selection section 19 has a document-retrieval means by the reference formula especially, the publication number by document-retrieval equipment can be received as an input, and the related keyword outputted can be made to reflect in the reference formula input section of document-retrieval equipment by considering as another composition as shown in later drawing 7 .
[0038] Thus, when the subset of the document which is a part of the target documents is specified according to the form of this operation, By extracting the part of the word groups which calculated significance, aligned in order of significance, and aligned about each of each word which appears in each document contained in the subset concerned, and considering as a related keyword It has the effect that it can ask for the related keyword based on the property of the target document dynamically and at high speed.
[0039] Moreover, it can use the related keyword obtained as mentioned above as an input to the document-retrieval equipment for the same document, and the exact keyword which suited the property of an object document in that case is not only reusable, but since surely being contained in an object document is guaranteed, the related keyword concerned has the effect that a reference result is surely obtained, when it searches using this.
[0040] The obtained related keyword can be used as an input to the document-retrieval equipment for the same object document set or another object document set. moreover, in that case In the document set set as the object of related keyword extraction based on a characteristic keyword The same or another document set can be searched and it has the effect that it is applicable also to the document set with a property which is different in the keyword concerned in the case of the document-retrieval equipment which makes another document set applicable to reference especially.
[0041] Moreover, it has the effect that it can use easily also in a user unfamiliar to operation of reference at the same time it becomes possible to choose a related keyword by simple operations, such as the click of a mouse, it mitigates the operation in re-reference and it raises the efficiency of reference instead of inputting reference conditions again from a keyboard by considering as the composition of making a user present and choose the extracted keyword, in case a user performs re- reference.
[0042] Moreover, if it is document-retrieval equipment which can give weight to each word in a reference condition, it has the effect that a highly precise reference result can obtain by considering the extracted keyword and its significance as an input as it is, in the document-retrieval equipment which calculates a goodness of fit for example, with reference conditions, and carries out the ranking of the document by adding and showing the extracted related keyword the significance.
[0043] Moreover, by accumulating the group of the extracted related keyword and its significance as a user's history, it becomes possible to express a range, taste, etc. of interest of a user as a vector of a keyword and its weight, and also has the effect that the large application of using this vector for reference of other document sets etc. is possible.
[0044] (Form 2 of operation) Next, the form of operation of the 2nd of this invention is explained using the same drawing 1 as the block diagram shown in the form 1 of operation. With the form of this 2nd operation, the

document set selection section 19 specifies two kinds of document sets A, and the document set B. Here, the document set B is the subset of the document set A. for example, the document set A obtained as a result of referring to a certain reference formula -- among those, they are the case where the document set B which the user specified as a related document group is specified, the case where the document set A specified according to the attribute of a document and the document set B further narrowed down by the reference formula in it are specified, etc.
[0045] The significance of a word is computed by carrying out the multiplication of the distribution index of the word computed by the following formulas in this case to the significance of the word concerned etc.
DI(A,B,W)= {(NA/DA(W))*(DB(W)/NB)}
However, the total number NB of documents of the number NA[of documents]:subset A in which the word "W" in the number DB[of documents] (W):subset B, in which the word "W" in the DA(W):subset A appears appears: The total number of documents of Subset B [0046] This appears by high frequency in Subset B, and serves as a value with what [higher] has the lower frequency of occurrence in Subset A. The word which serves as a high value in an upper formula contributes to the discrimination nature of Subset B greatly in Subset A, and it can be said that it is the keyword by which Subset B is characterized more. For example, in the example shown in drawing 3 , it supposes that the publication-number list 31 is Subset B, and suppose that it is the subset A containing this (it considers as the total 100 documents) as the number of appearance documents of each word in Subset A being the following by the case where it is specified simultaneously.
DA (applet) = 10DA(Internet) = 28DA (CGI) = 9DA (WWW) = 14DA (JAVA) = 20DA (SUN) = 5DA (script) = 10
[0047] In this case, the significance S2 of each word (W) serves as a value which carried out the multiplication of the weight DI (A, B, W) of each word to significance [of each word explained with the form 1 of operation] S (W), and is calculated as follows.
S2 (applet) = 92.096 * {(100/10)* (2/3)}
= 613.973 S2(Internet) = 57.258 * {(100/28)* (3/3)}
= 204.493 S2(CGI) = 23.024 * {(100/9)*(1/3)}
= 85.274 S2(WWW) = 18.945 * {(100/14)*(1/3)}
= 45.107 S2(JAVA) = 184.644 * {(100/20)*(3/3)}
= 923.220 S2(SUN) = 39.930 * {(100/5)*(1/3)}
= 266.200 S2 (script) = 17.550 * {(100/10)* (1/3)}
= It becomes the order of S2 (CGI) = 85.274 S2 (script) = 58.500 S2 (WWW) = 45.107. If it is set to 58.500 and aligns in order of significance S2 (JAVA) = 923.220 S2 (applet) = 613.973 S2 (SUN) = 266.200 S2(Internet) = 204.493 Therefore, if three high orders are extracted as a related keyword, "JAVA", an "applet", and "SUN" will serve as a related keyword.
[0048] The above-mentioned formula is an example and may use other formulas from which it appears by high frequency in Subset B, and what has the low frequency of occurrence in Subset A serves as a high value.
[0049] Thus, according to the form of this operation, it has the effect that a highly precise related keyword can be obtained, by taking into consideration the difference in the frequency distribution between two kinds of specified subsets.
[0050] (Form 3 of operation) Next, the form of operation of the 3rd of this invention is explained using the same drawing 1 as the block diagram shown in the form 1 of operation. With the form of this 3rd operation, the function which gives the weight of each document to the document set selection section 19 is prepared. For example, when a user specifies a document and five steps of evaluation values are given by making degree of association into an index to each document, or when ranking of the document obtained as a result of reference by the reference formula is carried out by the goodness of fit with a reference formula, it is the case where the weight which was said to the 1st place as ten points, and was said to the 2nd place as nine points is given etc. As opposed to the word contained in the document concerned, the multiplication of the weight given to each document is carried out, and the word ranking section considers it, and performs significance calculation. In addition, the weight given to each document may be a negative value. For example, in case a user specifies a document, a related document is also allowed weight grant of giving -one point to the document which is not related at all two points. Significance of the word included also in the document also irrelevant to a related document by this (and it is not so general) can be made low.
[0051] Thus, it has the effect that the highly precise related keyword which took the significance of each document into consideration is obtained by considering as a formula from which the word contained in a more important document serves as a high significance, by giving weight to each document contained in the specified

document set according to the gestalt of this operation.
[0052] (Gestalt 4 of operation) Next, the gestalt of operation of the 4th of this invention is explained. Drawing 4 is the block diagram of the related keyword automatic-extracting equipment concerning the gestalt of operation of the 4th of this invention. In addition to the 1st composition of the gestalt of operation, it has the threshold setting section 22, and changes, and transmission and reception of data have come to be able to do this threshold setting section among the word statistical information Management Department 17 with the gestalt of this 4th operation. Moreover, the word exclusion function by the threshold is given to the word statistical information Management Department 17 in the gestalt of this operation. In this composition, in case the word statistical information Management Department 17 outputs the statistical information of each word, with reference to the threshold setup 22 defined beforehand, the word of extremely high frequency or low frequency can consider it as the composition which excepts from a candidate on that spot and does not output the information on the word concerned to the word ranking section 18. For example, by setting up a threshold 1 with "the word which appears in 50% or more of a whole sentence document", and setting up a threshold 2 with "the word which appears only in one document", the bad influence which these words have on significance calculation can be prevented in advance, and improvement in the speed of processing can be attained.
[0053] In addition, according to characteristic quantity of the word concerned, such as the length of a word, you may set a threshold as several kinds in that case. For example, it is performing a threshold setup "the word of 50% or more of the whole and a single character being 30% or more of the whole for the word of two or more characters" by the case of Japanese, and the range of the word excepted in accordance with the property of each word is set up.
[0054] Thus, according to the form of this operation, by excepting the word whose appearance degree is high frequency or low frequency in the whole object document set in consideration of the threshold which was able to be defined beforehand, keyword extraction processing can be accelerated and it has the effect that only a word with high effectiveness can be sorted out in the case of reuse.
[0055] (Form 5 of operation) Next, the form of operation of the 5th of this invention is explained. Drawing 5 is the block diagram showing the composition of the related keyword automatic-extracting equipment concerning the form of operation of the 5th of this invention. The related keyword automatic-extracting equipment concerning the form of this 5th operation As [explained / in the form of the 1st operation] The word statistical information 14 of the whole document, and the word statistical information 15 in a document It adds to the basic composition which has the word statistical information Management Department 17 which manages, the word ranking section 18, the document set selection section 19 which specifies the subset of an object document, and the condition input section 20 which is a selection condition input means to the document set selection section 19. It aims at raising the quality of the related keyword group extracted by the word ranking section's 18 being interlocked with and using various information, such as the attribute of a word. In drawing 5 , as for the appearance Research and Data Processing Department and 26, the word attribute Research and Data Processing Department and 27 are the character string inclusion relation judging sections, these function parts are contained in related keyword automatic-extracting equipment 29, and a sign 25 is interlocked with the word ranking section. Moreover, 28 is the representation word grant section and this representation word grant section 28 outputs a related keyword in response to data from the word ranking section 18. Moreover, the word appearance positional information extraction section 23 which extracts the information on a position that a word appears based on the data from the object document set 11, as an external function part is formed to related keyword automatic-extracting equipment 29, and the appearance positional information 24 is outputted from this word appearance positional information extraction section 23. This appearance information is sent to the appearance Research and Data Processing Department 25.
[0056] The operation is explained about the gestalt of the operation of the 5th of this invention which has this composition. In operation of the gestalt of this operation, the statistical information extraction section 13 which operates as pretreatment to the target document set 11 using a dictionary 12 first extracts the word statistical information 14, such as the frequency of occurrence, a distribution, etc. of the word in the object document set 11 whole, and the word statistical information 15 in a document which is the statistical information of the word contained in the document concerned for every document. Simultaneously, if there is need, the word positional information extraction section 23 will also extract the appearance positional information 24 of a word. Drawing 6 is a table format view showing an example of the data structure of the appearance positional information 24 extracted by the word appearance positional information extraction section 23. Appearance positional information is stored as a table as shown in drawing 6 . The word which appears in the document for every

document, an appearance position (for example, byte offset from the head of a document), an appearance partition, etc. are stored.
[0057] And on the occasion of related keyword automatic-extracting operation, the appearance Research and Data Processing Department 25 is asked to each word, information, such as an appearance position of the word concerned and the appearance context, is acquired, and significance calculation is seasoned with this. For example, significance is computed by the technique of carrying out the multiplication of the "weight" like one point to the significance of each word, when are contained in a title by in any the word concerned is contained among these elements when all the documents made applicable to reference are documents which consist of elements, such as a title (or header), a subtitle, and the text, it is contained in a three-point subtitle by it and it be contained in the two-point text.
[0058] Or you may use the information on an appearance position. For example, it is also possible to compute significance by the technique of carrying out the multiplication of the "weight" like one point to the significance of the word concerned, if the number of characters between the word contained in a reference formula when the word contained in this reference formula can be referred to by the case where a subset is specified by the reference formula, and the word set as the object of the present significance calculation is less than two characters, it is three characters [less than ten] and it is two characters.
[0059] Moreover, as another mode of the form of this operation, to each word, the word attribute Research and Data Processing Department 26 is asked, and the part of speech of the word concerned, a classification, etc. acquire the attribute of the word, and season significance calculation with this. For example, it is also possible to compute significance paying attention to the part of speech of the word concerned, by the technique of in addition to this not being an independent word one point of carrying out the multiplication of the "weight" like zero point to the significance of each word if it becomes things (a particle, auxiliary verb, etc.), if it is a proper noun, it is a five-point common noun, it is a four-point adjective and an adjective verb and it is a two-point verb and an adverb.
[0060] Moreover, it judges whether inclusion relation is between one word in the words extracted as another mode of the gestalt of this operation using the character string inclusion relation judging section 27 which judges the inclusion relation as a character string between some two words, or the word group specified beforehand, and the extracted word, and the word to extract is restricted when judged with there being inclusion relation. The word group specified beforehand here is a word contained in the reference formula at the time of using a reference formula for specification of a subset. In the judgment of inclusion relation, the case where any one of the following criteria (one [or] or more) is filled can be recognized as inclusion relation by setup defined beforehand.
The word "A", and the word "B" are in agreement in the front. the word "A" (1) When shorter than the word "B", (2) The word "A", and the word "B" are in agreement in back. the word "A" When shorter than the word "B", (3) It is [0061] when completely in agreement [the word "A" is the portion of the word "B", and when the front and back are not in agreement, the relation between (4) words "A", and the word "B" fills either of (1) - (3), and] with the component of the word "B." For example, on the criteria of (1), "Tokyo" to "Tokyo" is judged to be a code word. Hereafter, on the criteria of (2), "gratitude" of as opposed to "large Thanksgiving Day" in "sale" to "new sale" is similarly judged by the criteria of (3), respectively to be a code word. If the criteria of (4) are important in the case of the code-word judging in English and these criteria are followed "artificial intelligence" It receives. "art" "tell" A code word is "artificail" although it does not become. "intelligence" It is judged with a code word.
[0062] One of the following criteria (set up beforehand) is followed [words / judged that have a code-word relation by the above-mentioned criteria / two] also about which / the / is adopted as a related keyword. (1) Adopt the word of a long unit (2). The word of a short unit is adopted (3). A word with a high significance is adopted (4). [0063] which adopts the difference of the word of a short unit and the word of a long unit, and the word of a short unit For example, a word "Tokyo" is extracted with significance 10 and a word "Tokyo" is extracted with significance 7, respectively. And when a code-word relation is materialized to both, if the criteria of (1) are followed, "Tokyo" long as a character string will be adopted, if the criteria of (2) are followed, "Tokyo" short as a character string will be adopted, and when the criteria of (3) are followed, "Tokyo" where significance is more high will be adopted. the criteria of (4) -- for example, word "artificial intelligence" "artificial" the case where a code-word relation is materialized in between -- "artificial" and -- "intelligence" It adopts as a related keyword and is mainly effective in an English document.
[0064] When it is the word in which a code-word relation is materialized between the word groups specified

beforehand, technique other than (3) can be used. In this case, it becomes the processing "it will not adopt as a related keyword if it is a short unit (or long unit)." Any technique can be used when a code-word relation is materialized in the extracted words.

[0065] Moreover, the extracted related keyword group is classified and shown as another mode of the form of this operation using the attribute and statistical information of each word. if a part of speech is used as an attribute of a word, other than this, a proper noun can be resembled, for example, and it can classify, and can show Or it is also possible to use a thesaurus dictionary as an attribute of a word, to classify each word according to the form corresponding to the classification in a thesaurus, and to show it. Moreover, with the classification using statistical information, the technique classified according to the number of appearance documents of each word in the document set specified, for example is raised. The effect of narrowing down at the time of the word being used for re-reference can be checked in advance by classifying according to the criteria "whether the number of appearance documents is 80 percent or more of a document set" in that case. In addition, when using a thesaurus dictionary as an attribute of a word in a classification, it is also possible to give the word equivalent to the host node of a thesaurus as a representation word to the classified word group, and to represent a word group with the word. Similarly, when using the statistical information 14 of a word, in the classified word group, you may adopt a word with the highest frequency of occurrence as a representation word.

[0066] Thus, according to the gestalt of this operation, the related keyword in consideration of the information on the structure of a document or the distance between words can be extracted by using the information on a position that the word appeared, and it has the effect that highly precise related keyword extraction becomes possible.

[0067] Moreover, the part of speech of a word etc. can extract the related keyword according to the feature of each attribute by taking into consideration the attribute information on each word, and it has the effect that highly precise related keyword extraction becomes possible.

[0068] Moreover, by taking into consideration the inclusion relation as a character string between words, the word which are the same meaning and a use is eliminated, a related keyword can be extracted, and it has the effect that the redundancy as the whole related keyword can be suppressed.

[0069] Moreover, the extracted related keyword is classified, by setting up the representation word corresponding to each classification, if there is need, the list nature of the extracted keyword, an inclination, the effectiveness in the case of reuse, etc. are checked beforehand, a related keyword can be extracted, and it has the effect that the facility as a related keyword can be improved.

[0070] (Form 6 of operation) Next, the form of operation of the 6th of this invention is explained. Drawing 7 is the block diagram showing the document-retrieval structure of a system realized combining the composition of document-retrieval equipment and this concerning the form of operation of the 6th of this invention, and related keyword automatic-extracting equipment. This document-retrieval equipment 41 cooperates with the related keyword automatic-extracting equipment concerning the form of the above 1st, the 2nd, the 3rd, the 4th, or the 5th operation, and operates.

[0071] The document-retrieval equipment 41 in this operation form has the document ranking section 46 which calculates the goodness of fit between the reference condition input section 44 which inputs conditional expression required for a document retrieval, the document-retrieval section 45 which searches a document according to the inputted reference conditions, and the reference formula and document which were inputted about the document searched in the document-retrieval section 45, and changes. This document-retrieval equipment 41 makes applicable to reference the same object document set 11 as the related keyword automatic-extracting equipment 48 which cooperates and operates, and searches using the index 43 for document retrievals beforehand created by the index generation section 42 using the same dictionary 12 as using for word statistical information extraction. Moreover, the related keyword automatic-extracting equipment 48 in this operation form considers the document set selection section 47 as another composition, and sets (list of publication numbers which are meaning etc.) of the identifier of the document corresponding to each element of the subset of the document which the document set selection section 47 specified are inputted into related keyword automatic-extracting equipment 48.

[0072] The operation is explained about the form of this operation equipped with the above composition. Based on the reference conditions first inputted into the reference condition input section 44, the document with which the document-retrieval section 45 suits reference conditions with reference to the index 43 for reference is specified. You may make what calculated the goodness of fit between the reference formulas and

documents which were inputted [in / the document ranking section 46 / further] although it was good also as a reference result document 50 as it is in the document set obtained here, and aligned the document in order with a high goodness of fit the composition of considering as a reference result. In this way, the document set 50 of the obtained reference result is passed to the document set selection section 47 as soon as it returns to a user as a reference result. In the document selection section 47, all or a part of document set passed from the document ranking section 46 is adopted as an input to related keyword automatic-extracting equipment 48. As long as ranking of the document is carried out to the order of a goodness of fit, you may make it the composition of selecting high order 10 document among document sets of a reference result. Moreover, if the attribute information beforehand given for every document can be used, it is good also as composition of selecting only the document which has specific attribute value using this.

[0073] The subset of the document specified by the document set selection section 47 is sent to related keyword automatic-extracting equipment 48, and extracts the related keyword group 49 in a procedure as shown in the form of the above 1st, the 2nd, the 3rd, the 4th, or the 5th operation. In this way, the obtained related keyword group 49 is returned to the reference condition input section 44, and a user is shown it. A user can choose a required thing from the shown related keyword group, can consider as new reference conditions, and can perform reference again. According to the form of this operation, the related keyword obtained as mentioned above by related keyword automatic-extracting equipment by this It can use as an input to the document-retrieval equipment for the same document. In this case, the exact keyword which suited the property of an object document is not only reusable, but since surely being contained in an object document is guaranteed, the related keyword concerned has the effect that a reference result is surely obtained, when it searches using this.

[0074] (Gestalt 7 of operation) Next, the gestalt of operation of the 7th of this invention is explained. Drawing 8 is the block diagram showing the document-retrieval structure of a system realized combining the composition of document-retrieval equipment and this concerning the gestalt of operation of the 7th of this invention, and related keyword automatic-extracting equipment. This document-retrieval equipment 51 cooperates with the document-retrieval equipment 41 concerning the gestalt of the 6th operation, and the related keyword automatic-extracting equipment similarly applied to the gestalt of the above 1st, the 2nd, the 3rd, the 4th, or the 5th operation, and operates.

[0075] The document-retrieval equipment 51 in this operation form has the reference condition input section 54 which inputs conditional expression required for a document retrieval, and the document-retrieval section 55 which searches a document according to the inputted reference conditions, and changes. The document-retrieval equipment 51 in this operation form makes applicable to reference object document set 56 which is different in the related keyword automatic-extracting equipment 52 which cooperates and operates, and has the composition that the document-retrieval section 55 is connected to the object document set 56. In addition, the detail about the reference technique is not asked.

[0076] Operation in this operation form equipped with the above composition is explained below. According to the conditions specified first, related keyword automatic-extracting equipment 52 operates, and the related keyword group 53 is outputted. The reference condition input section 54 in document-retrieval equipment 51 is shown to a user by considering the related keyword group 53 as an input, and a user can choose only a required thing among the shown related keywords, and can perform reference to the object document set 56 used as the candidate for reference, and it can obtain the reference result document 57.

[0077] Thus, according to the form of this operation, the related keyword obtained by related keyword automatic-extracting equipment 52 can be used as an input to the document-retrieval equipment 51 for the same object document set or another object document set. In the document set set as the object of related keyword extraction in this case, based on a characteristic keyword The same or another document set can be searched and it has the effect that it is applicable also to the document set with a property which is different in the keyword concerned in the case of the document-retrieval equipment which makes another document set applicable to reference especially.

[0078]

[Effect of the Invention] The document set selection section which specifies the subset of a document for related keyword automatic-extracting equipment according to this invention as explained above, the word which appears for the whole object document or each document of every, the word statistical information Management Department which manages the statistical information, and the word ranking section which computes the significance of each word which appears in the subset of a document, and aligns in order of

significance -- **, since it was alike and constituted more It is possible to ask for the whole document and the statistical information of each word in the specified document subset at high speed, ranking of each word which appears in the specified document set can be carried out at high speed based on the significance, and the part can be shown as a related keyword.
[0079] Moreover, the precision as a related term of the word group extracted can be raised by deleting the word which the weight of a word is changed or fulfills specific conditions from the word group after ranking by establishing a means to manage the appearance position in the attribute information on a word, or a document etc. in addition to the aforementioned composition. Moreover, more intelligible related keyword presentation can be performed by classifying the extracted word group according to the attribute and statistical property of a word.
[0080] Furthermore, by constituting the document-retrieval system containing the document-retrieval equipment which cooperated with related keyword automatic-extracting equipment, and reusing the extracted related keyword as an input If the extracted related keyword suits the property of an object document and the candidate for reference is the same document group, since it is secured that at least one or more reference results are obtained by the keyword, The effect of being able to perform re-reference efficiently and easily is acquired.

[Translation done.]

(19)日本国特許庁（ＪＰ）　(12)公開特許公報（Ａ）　(11)特許出願公開番号

特開平11－25108

(43)公開日　平成11年(1999)１月29日

(51)Int.Cl.[6]	識別記号	ＦＩ	
Ｇ０６Ｆ　17/30		Ｇ０６Ｆ　15/401	３１０Ａ
		15/40	３７０Ａ

審査請求　未請求　請求項の数17　ＯＬ　（全 17 頁）

(21)出願番号　特願平9－176822

(22)出願日　平成９年(1997)７月２日

(71)出願人　000005821
松下電器産業株式会社
大阪府門真市大字門真1006番地

(72)発明者　佐　藤　光　弘
大阪府門真市大字門真1006番地　松下電器産業株式会社内

(72)発明者　野　口　直　彦
大阪府門真市大字門真1006番地　松下電器産業株式会社内

(72)発明者　菅　野　祐　司
大阪府門真市大字門真1006番地　松下電器産業株式会社内

(74)代理人　弁理士　蔵合　正博

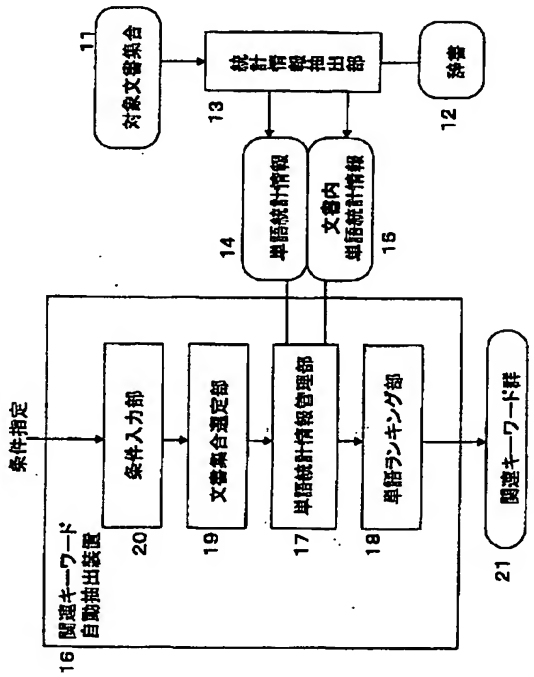
最終頁に続く

(54)【発明の名称】　関連キーワード自動抽出装置、文書検索装置及びこれらを用いた文書検索システム

(57)【要約】

【課題】　実際の検索対象文書の特性に即し、かつそのキーワードによる検索を実行した場合少なくとも１件以上の検索結果が得られるような関連キーワードを自動抽出すること。

【解決手段】　関連キーワード自動抽出装置として、各文書の属性情報や入力検索式などに基づいて文書の部分集合を特定する文書集合選定部１９と、各単語の対象文書１１全体における統計情報１４および文書毎に出現する単語とその統計情報１５を管理する単語統計情報管理部１７と、単語統計情報１４、１５を基に、或る文書の部分集合に出現する各単語の重要度を算出して重要度の順に整列する単語ランキング部１８とを設け、単語統計情報管理部により、文書全体、および特定された文書部分集合における各単語の統計情報を高速に求める。これにより、或る文書集合に出現する単語を、その重要度に基づいてランキングし、その一部を関連キーワードとして提示することができる。



1

【特許請求の範囲】

【請求項１】　辞書を用いて対象文書集合の各文書に出現する単語または単語の組の出現頻度や分布などの統計情報があらかじめ抽出されている文書集合に対して、各文書に付与された属性情報やユーザが入力した検索式などに基づいて文書の部分集合を特定する文書集合選定部と、各単語の対象文書全体における統計情報、および各文書ごとの当該文書に出現する単語とその統計情報を管理する単語統計情報管理部と、各単語の全文書および各文書ごとの統計情報を基に、特定された部分集合に出現する各単語の重要度を算出して重要度の順に整列する単語ランキング部とを有し、整列された単語群のうちの特定部分のみについて、単語もしくは単語とその重要度の組を抽出し、これを再利用可能な形で提示するようにしたことを特徴とする関連キーワード自動抽出装置。

【請求項２】　前記構成において、特定された部分集合Ａに対して、これに含まれる部分集合Ｂが文書集合選定部により特定された場合に、部分集合Ａに含まれる文書群に出現する単語の統計情報と、部分集合Ｂに含まれる文書群に出現する単語の統計情報との差分を、部分集合Ｂにおける各単語の重要度に加味することで、部分集合Ｂに出現する各単語の重要度を算出して単語ランキングに反映することを特徴とする請求項１に記載の関連キーワード自動抽出装置。

【請求項３】　文書集合選定部に各文書の重みを付与する機能を設け、特定された文書集合の各文書に含まれる単語の重要度に当該文書の重みを加味することにより当該単語の重要度を算出して単語ランキングに反映することを特徴とする請求項１または２に記載の関連キーワード自動抽出装置。

【請求項４】　対象文書集合全体において出現度合いが高頻度または低頻度である単語をあらかじめ定められた閾値を考慮して除外することにより、再利用の際に有効性の高い単語のみが選別できることを特徴とする請求項１乃至３のいずれかに記載の関連キーワード自動抽出装置。

【請求項５】　単語の長さなどその単語の特徴量に応じて除外のための閾値を変化させることにより再利用の際に有効性の高い単語のみが選別できることを特徴とする請求項４に記載の関連キーワード自動抽出装置。

【請求項６】　単語の出現位置や出現する文脈の情報を管理する出現情報管理部を有し、単語の重要度にその単語の出現情報の種類に応じてあらかじめ定められた重みを加味することにより当該単語の重要度を算出して単語ランキングに反映することを特徴とする請求項１乃至５のいずれかに記載の関連キーワード自動抽出装置。

【請求項７】　単語の品詞など、各単語の属性情報を管理する言語属性管理部を有し、当該単語の属性に応じてあらかじめ定められた重みを加味することにより当該単語の重要度を算出して単語ランキングに反映することを特徴とする請求項１乃至６のいずれかに記載の関連キーワード自動抽出装置。

2

【請求項８】　抽出された単語同士、またはあらかじめ指定された単語群と抽出された単語との間の文字列としての包含関係を、定められた条件により判定する文字列包含関係判定部を有し、当該単語同士に文字列としての包含関係があると判定された場合に、指定された条件に従って、長単位の文字列のみ、もしくは短単位の文字列のみ、もしくは重要度の高い方の文字列のみ、もしくは短単位の文字列および長単位の文字列と短単位の文字列との差分の双方、のいずれかを選択することにより、再利用の際に有効性の高い単語のみが選別できることを特徴とする請求項１乃至７のいずれかに記載の関連キーワード自動抽出装置。

【請求項９】　単語の品詞など、各単語の属性情報を管理する言語属性管理部を有し、当該単語の属性や、指定された部分集合または文書全体における出現頻度、分布等を考慮することにより、抽出された単語を分類して提示できることを特徴とする請求項１乃至８のいずれかに記載の関連キーワード自動抽出装置。

【請求項１０】　分類された単語群のそれぞれについて、その集合を代表する単語を付与する代表語付与部を設け、分類された単語群を代表する代表語群のみ、もしくは代表語と全ての単語を提示できることを特徴とする請求項９に記載の関連キーワード自動抽出装置。

【請求項１１】　辞書を用いて対象文書集合の各文書に出現する単語または単語の組の出現頻度や分布などの統計情報があらかじめ抽出されている文書集合に対して、文書検索に必要な条件式を入力する検索条件入力部と、入力された検索条件にしたがって対象文書集合から文書の検索を行なう文書検索部と、文書検索部４５において検索された文書について、入力された検索式と文書との間の適合度を計算する文書ランキング部４６とを有して成り、文書ランキング部におけるランキング結果を関連キーワード自動抽出装置へ送付し、また関連キーワード自動抽出装置からフィードバックされた関連キーワードを検索条件入力部へ入力することが可能な文書検索装置。

【請求項１２】　文書検索に必要な条件式を入力する検索条件入力部と、入力された検索条件にしたがって対象文書集合から文書の検索を行なう文書検索部とを有して成り、前記検索条件入力部は、ユーザが検索条件を入力するのに加えて関連キーワード自動抽出装置から送られてきた関連キーワードを検索条件として入力することが可能な文書検索装置。

【請求項１３】　辞書を用いて対象文書集合の各文書に出現する単語または単語の組の出現頻度や分布などの統計情報があらかじめ抽出されている文書集合に対して、文書検索に必要な条件式を入力する検索条件入力部と、入力された検索条件にしたがって対象文書集合から文書

の検索を行なう文書検索部と、文書検索部４５において検索された文書について、入力された検索式と文書との間の適合度を計算する文書ランキング部４６とを有して成る文書検索装置と、
前記文書検索装置に接続された関連キーワード自動抽出装置とから構成され、
前記文書検索装置の文書ランキング部から出力されたランキング結果を関連キーワード自動抽出装置へ送付し、また関連キーワード自動抽出装置から文書検索装置の検索条件入力部へ関連キーワードをフィードバックしてキーワード検索を行なうようにしたことを特徴とする文書検索システム。

【請求項１４】　文書検索装置と関連キーワード自動抽出装置との間には文書集合選定部が設けられ、文書検索装置の文書ランキング部から出力されたランキング結果は文書集合選定部に送付されて文書の特定が行なわれ、前記関連キーワード自動抽出装置４８には、文書集合選定部４７が特定した文書の部分集合が入力されることを特徴とする請求項１３記載の文書検索システム。

【請求項１５】　関連キーワード自動抽出装置には、請求項１乃至１０のいずれかに記載の関連キーワード自動抽出装置が用いられることを特徴とする請求項１３または１４記載の文書検索システム。

【請求項１６】　文書検索に必要な条件式を入力する検索条件入力部と、入力された検索条件にしたがって対象文書集合から文書の検索を行なう文書検索部とを有して成る文書検索装置と、
前記文書検索装置に接続された関連キーワード自動抽出装置とから構成され、
前記文書検索装置の検索条件入力部は、ユーザが検索条件を入力するのに加えて関連キーワード自動抽出装置から送られてきた関連キーワードを検索条件として入力してキーワード検索を行なうようにしたことを特徴とする文書検索システム。

【請求項１７】　関連キーワード自動抽出装置には、請求項１乃至１０のいずれかに記載の関連キーワード自動抽出装置が用いられることを特徴とする請求項１６記載の文書検索システム。

【発明の詳細な説明】

【０００１】

【産業上の利用分野】本発明は、特定の文書集合から、その文書集合を特徴づける語句をキーワードとして抽出するための関連キーワード自動抽出装置、および前記関連キーワード自動抽出装置を利用した文書検索装置に関する。

【０００２】

【従来の技術】文書検索装置において、ユーザが必要とする文書を得るためには、適切な検索語を利用した検索式を入力する必要があるが、ユーザ自身が適切な検索語を想起し難い、という問題がある。そこで従来、ユーザが入力した検索語に対して、関連語辞書などを利用して検索語に関連する語を提示することにより、ユーザの再検索を助ける手法などが取られてきた。しかしながら、こうした手法はあらかじめ静的にさだめられた関連語辞書の性質に依存するため、検索対象となる文書の特性に即した関連語が得られない。また、得られた単語で検索した結果少なくとも１件以上の文書が得られることが保証されない、という欠点があった。

【０００３】

【発明が解決しようとする課題】本発明は前記の課題を解決するもので、特定された文書集合における各単語の出現頻度・分布などの統計情報と、検索対象文書全体における単語の統計情報とを考慮して単語の重要度を算出し、これにもとづいて単語をその重要度によってランキングし、ランクの一部である単語群を抽出することにより、実際の検索対象文書の特性に即し、かつ品質の高い関連キーワード群を高速かつ動的に抽出できる、関連キーワード自動抽出装置を提供することを目的とする。

【０００４】また、前記関連キーワード自動抽出装置から得られた関連キーワード群を利用して検索を実行した場合、少なくとも１件以上の検索結果が得られることを保証する文書検索装置及びこれらを用いた文書検索システムを提供することを目的とするものである。

【０００５】

【課題を解決するための手段】本発明は、上記目的を達成するため、関連キーワード自動抽出装置として、各文書に付与された属性情報やユーザが入力した検索式などに基づいて文書の部分集合を特定する文書集合選定部と、各単語の対象文書全体における統計情報や各文書ごとに出現する単語とその統計情報を管理する単語統計情報管理部と、各単語の全文書または各文書内統計情報を基に、特定された文書の部分集合に出現する各単語の重要度を算出して重要度の順に整列する単語ランキング部とを設け、単語統計情報管理部により、文書全体、および特定された文書部分集合における各単語の統計情報を高速に求めることが可能であり、特定された文書集合に出現する各単語を、その重要度の順に高速にランキングし、その一部を関連キーワードとして提示することができる。

【０００６】さらに、前記構成に加えて、単語の属性情報や文書内の出現位置を管理する手段などを設けることにより、単語の重みを変化させ、あるいはランキング後の単語群から特定の条件を満たす単語を削除することで、抽出される単語群の関連語としての精度を向上させることができ、また、抽出された単語群を、語の属性や統計的性質により分類することで、よりわかりやすい関連キーワード提示を行なうことができる。

【０００７】また本発明は、上記目的を達成するため、関連キーワード自動抽出装置と連携した文書検索装置を含む文書検索システムを構成し、抽出された関連キーワ

ードを入力として再利用することにより、抽出された関連キーワードが対象文書の特性に合ったものであり、かつ検索対象が同一の文書群であるならば、そのキーワードによって検索結果が少なくとも１件以上得られることが保障されるため、効率的かつ容易に再検索を行なうことができる。

【０００８】

【発明の実施の形態】本発明の請求項１に記載の発明は、辞書を用いて対象文書集合の各文書に出現する単語または単語の組の出現頻度や分布などの統計情報があらかじめ抽出されている文書集合に対して、各文書に付与された属性情報やユーザが入力した検索式などに基づいて文書の部分集合を特定する文書集合選定部と、各単語の対象文書全体における統計情報、および各文書ごとの当該文書に出現する単語とその統計情報を管理する単語統計情報管理部と、各単語の全文書および各文書ごとの統計情報を基に、特定された部分集合に出現する各単語の重要度を算出して重要度の順に整列する単語ランキング部とを備えたものであり、整列された単語群のうちの特定部分のみについて、単語もしくは単語とその重要度の組を抽出し、これを再利用可能な形で高速提示するという作用を有する。

【０００９】本発明の請求項２に記載の発明は、請求項１に記載の関連キーワード自動抽出装置において、特定された部分集合Ａに対して、これに含まれる部分集合Ｂが文書集合選定部により特定された場合に、部分集合Ａに含まれる文書群に出現する単語の統計情報と、部分集合Ｂに含まれる文書群に出現する単語の統計情報との差分を、部分集合Ｂにおける各単語の重要度に加味することで、部分集合Ｂに出現する各単語の重要度を算出して単語ランキングに反映するようにしたものである。

【００１０】本発明の請求項３に記載の発明は、請求項１または２に記載の関連キーワード自動抽出装置において、文書集合選定部に各文書の重みを付与する機能を設け、特定された文書集合の各文書に含まれる単語の重要度に当該文書の重みを加味することにより当該単語の重要度を算出して単語ランキングに反映するようにしたものである。

【００１１】本発明の請求項４に記載の発明は、請求項１乃至３のいずれかに記載の関連キーワード自動抽出装置において、対象文書集合全体において出現度合いが高頻度または低頻度である単語をあらかじめ定められた閾値を考慮して関連キーワード抽出の対象から除外することにより、再利用の際に有効性の高い単語のみが選別できるようにしたものである。

【００１２】本発明の請求項５に記載の発明は、請求項４に記載の関連キーワード自動抽出装置において、単語の長さなどその単語の特徴量に応じて除外のための閾値を変化させることにより再利用の際に有効性の高い単語のみが選別できるようにしたものである。

【００１３】本発明の請求項６に記載の発明は、請求項１乃至５のいずれかに記載の関連キーワード自動抽出装置において、単語の出現位置や出現する文脈の情報を管理する出現情報管理部を有し、単語の重要度にその単語の出現情報の種類に応じてあらかじめ定められた重みを加味することにより当該単語の重要度を算出して単語ランキングに反映するようにしたものである。

【００１４】本発明の請求項７に記載の発明は、請求項１乃至６のいずれかに記載の関連キーワード自動抽出装置において、単語の品詞など、各単語の属性情報を管理する言語属性管理部を有し、当該単語の属性に応じてあらかじめ定められた重みを加味することにより当該単語の重要度を算出して単語ランキングに反映するようにしたものである。

【００１５】本発明の請求項８に記載の発明は、請求項１乃至７のいずれかに記載の関連キーワード自動抽出装置において、抽出された単語同士、またはあらかじめ指定された単語群と抽出された単語との間の文字列としての包含関係を、定められた条件により判定する文字列包含関係判定部を有し、当該単語同士に文字列としての包含関係があると判定された場合に、指定された条件に従って、長単位の文字列のみ、もしくは短単位の文字列のみ、もしくは重要度の高い方の文字列のみ、もしくは短単位の文字列および長単位の文字列と短単位の文字列との差分の双方、のいずれかを選択することにより、再利用の際に有効性の高い単語のみが選別できるようにしたものである。

【００１６】本発明の請求項９に記載の発明は、請求項１乃至８のいずれかに記載の関連キーワード自動抽出装置において、単語の品詞など、各単語の属性情報を管理する言語属性管理部を有し、当該単語の属性や、指定された部分集合または文書全体における出現頻度、分布等を考慮することにより、抽出された単語を分類して提示できるようにしたものである。

【００１７】本発明の請求項１０に記載の発明は、請求項９に記載の関連キーワード自動抽出装置において、分類された単語群のそれぞれについて、その集合を代表する単語を付与する代表語付与部を設け、分類された単語群を代表する代表語群のみ、もしくは代表語と全ての単語を提示できるようにしたものである。

【００１８】本発明の請求項１１に記載の発明は、文書検索装置として、辞書を用いて対象文書集合の各文書に出現する単語または単語の組の出現頻度や分布などの統計情報があらかじめ抽出されている文書集合に対して、文書検索に必要な条件式を入力する検索条件入力部と、入力された検索条件にしたがって対象文書集合から文書の検索を行なう文書検索部と、文書検索部において検索された文書について、入力された検索式と文書との間の適合度を計算する文書ランキング部とを備えたものであり、文書ランキング部におけるランキング結果を関連キ

ーワード自動抽出装置へ送付し、また関連キーワード自動抽出装置からフィードバックされた関連キーワードを検索条件入力部へ入力するという作用を有する。

【００１９】本発明の請求項１２に記載の発明は、文書検索装置として、文書検索に必要な条件式を入力する検索条件入力部と、入力された検索条件にしたがって対象文書集合から文書の検索を行なう文書検索部とを備えたものであり、前記検索条件入力部は、ユーザが検索条件を入力するのに加えて関連キーワード自動抽出装置から送られてきた関連キーワードを検索条件として入力するという作用を有する。

【００２０】本発明の請求項１３に記載の発明は、文書検索システムとして、辞書を用いて対象文書集合の各文書に出現する単語または単語の組の出現頻度や分布などの統計情報があらかじめ抽出されている文書集合に対して、文書検索に必要な条件式を入力する検索条件入力部と、入力された検索条件にしたがって対象文書集合から文書の検索を行なう文書検索部と、文書検索部において検索された文書について、入力された検索式と文書との間の適合度を計算する文書ランキング部とを有して成る文書検索装置と、前記文書検索装置に接続された関連キーワード自動抽出装置とを備えたものであり、前記文書検索装置の文書ランキング部から出力されたランキング結果を関連キーワード自動抽出装置へ送付し、また関連キーワード自動抽出装置から文書検索装置の検索条件入力部へ関連キーワードをフィードバックしてキーワード検索を行なうという作用を有する。

【００２１】本発明の請求項１４に記載の発明は、請求項１３記載の文書検索システムにおいて、文書検索装置と関連キーワード自動抽出装置との間には文書集合選定部が設けられ、文書検索装置の文書ランキング部から出力されたランキング結果は文書集合選定部に送付されて文書の特定が行なわれ、前記関連キーワード自動抽出装置４８には、文書集合選定部４７が特定した文書の部分集合が入力されるようにしたものである。

【００２２】本発明の請求項１５に記載の発明は、請求項１３または１４記載の文書検索システムにおいて、関連キーワード自動抽出装置には、請求項１乃至１０のいずれかに記載の関連キーワード自動抽出装置が用いられるようにしたものである。

【００２３】本発明の請求項１６に記載の発明は、文書検索に必要な条件式を入力する検索条件入力部と、入力された検索条件にしたがって対象文書集合から文書の検索を行なう文書検索部とを有して成る文書検索装置と、前記文書検索装置に接続された関連キーワード自動抽出装置とを備えたものであり、前記文書検索装置の検索条件入力部は、ユーザが検索条件を入力するのに加えて関連キーワード自動抽出装置から送られてきた関連キーワードを検索条件として入力してキーワード検索を行なうという作用を有する。

【００２４】本発明の請求項１７に記載の発明は、請求項１６記載の文書検索システムにおいて、関連キーワード自動抽出装置には、請求項１乃至１０のいずれかに記載の関連キーワード自動抽出装置が用いられるようにしたものである。

【００２５】以下に、本発明の具体的な実施の形態について、添付の図面を参照して説明する。

【００２６】（実施の形態１）最初に、本発明の第１の実施の形態について説明する。図１は本発明の第１の実施の形態に係る関連キーワード自動抽出装置の構成を示したブロック図である。まず、対象となる文書集合１１に対し、辞書１２を利用して、前処理として動作する統計情報抽出部１３により、文書集合全体における単語の頻度・分布などの単語統計情報１４、および各文書ごとの当該文書に含まれる単語の統計情報である文書内単語統計情報１５を抽出しておく。図２（ａ）は単語統計情報の構造を示すテーブル構成図であり、図２（ｂ）は文書内単語統計情報の構造を示すテーブル構成図である。単語統計情報１４は、統計情報抽出部１３によって抽出された単語の統計情報を例えば図２（ａ）に示すようなテーブルとして格納する。このテーブルを利用することにより、例えば単語「インターネット」の全文書中での総出現頻度や出現文書数を高速に求めることができる。また、文書内単語統計情報１５は各文書ごとの単語の統計情報を例えば図２（ｂ）に示すようなテーブルとして格納する。これにより、例えば文書番号００１０には単語「インターネット」が５回、単語「WWW」が２回出現する、といった各文書ごとの統計情報を高速に求めることができる。

【００２７】関連キーワード自動抽出装置１６は、文書全体の単語統計情報１４および文書内単語統計情報１５を管理する単語統計情報管理部１７と、単語の重要度を算出する単語ランキング部１８と、対象文書の部分集合を特定する文書集合選定部１９と、文書集合選定部１９への選定条件を入力する手段である条件入力部２０とから構成される。

【００２８】かかる構成を有する関連キーワード自動抽出装置１６の動作について以下説明する。最初に、条件入力部２０に対して入力された条件により、文書集合選定部１９が文書集合を特定する。文書集合は、次の３種類の手段のいずれかまたはその組み合わせにより特定される。

（１）文書の属性により文書集合を特定する。この場合、文書集合選定部１９は文書の所属するジャンルなど、文書にあらかじめ付与された属性値により文書を選択する手段を有し、条件入力部２０により指定された属性値に合致する文書群を部分集合として採用する。

（２）検索式により文書集合を特定する。この場合、文書集合選定部１９は条件入力部２０で入力された検索式に適合する文書を特定する文書検索手段を有し、これを

利用して検索の結果得られる文書群を部分集合として採用する。なおその際、文書検索手段に検索式との適合度を判定して文書を適合度の順にランキングする機能があるならば、検索結果のうちの特定部分、例えば上位１０文書を部分集合として採用しても良い。

（３）ユーザにより指定された文書集合。この場合、文書集合選定部１９は条件入力部２０においてユーザが直接指定した（複数の）文書を部分集合として採用する。

【００２９】文書集合選定部１９は、以上により選定された文書集合を各文書を一意に決定する識別子の集合、例えば文書番号のリストとして単語統計情報管理部１７に渡す。単語統計情報管理部１７は、特定された文書集合に対して、文書ごとに文書番号から文書内単語統計情報１４を調べ、当該文書に出現する単語とそれぞれの文書内の出現頻度を得る。次に得られた単語すべてについて単語統計情報１５を調べ、当該単語の全文書における頻度や分布情報を得る。

【００３０】ここで得られた各種統計情報は単語ランキング部１８に渡され、各単語の重要度が算出される。ある単語Wの重要度Ｓ（W）は、例えば次のようにして算出することができる。

【数１】

$$S(W) = C * \sum_{j=0}^{n} \{TF_j(W) * IDF(W)\} * FN(W)$$

ただし

Ｃ　　：定数

ｎ　　：特定された文書集合に含まれる文書数

ＴＦｊ（W）：文書Ｄｊ における単語Wの出現頻度

ＦＮ（W）　：特定された文書集合中で単語Wを含む文書数

である。

【００３１】またＩＤＦ（W）は、単語Wのｉｄｆ値と呼ばれる指標であり、例えば以下の式により計算される。

ＩＤＦ（W）＝　１－ｌｏｇ（ＤＦ（W）／Ｎ）

ただし、

ＤＦ（W）：文書全体において単語Wが出現する文書数

Ｎ　　：全文書数

である。

【００３２】ＩＤＦ（W）は、単語Wがより多くの文書に出現する（すなわちより一般的な語である）場合にその値が小さくなる。これにより、対象文書全体において比較的よく出現する語の重要度を低く抑えることができる。さらにＦＮ（W）を考慮することで、特定された文書集合に多く現れる単語の重要度を高くでき、結果その特定文書集合に特徴的な語に高い重要度を与えることができる。なお、上記算出法において、ＴＦ（W）をその単語が含まれる文書の文書サイズ（文字数や含まれる単語の異なり数など）や単語の総出現頻度などで正規化してもよい。

【００３３】単語ランキング部１８は、特定された部分集合中の全文書に含まれる全単語について重要度計算を行い、その後全単語を重要度の順に整列する。最後に、整列された単語群から特定部分、例えば上位１０単語を採用し、単語、もしくは単語とその重要度の組として提示する。なお、抽出の際に重要度だけでなく、重要度算出に利用した各種統計情報などを同時に提示してもよい。また、抽出された関連キーワードとその重要度の組を、例えばユーザの履歴として蓄積していくこともできる。このようにすることにより、ユーザの興味の範囲や嗜好などをキーワードとその重みのベクトルとして表現することが可能となり、このベクトルを他の操作、例えば文書集合の検索に利用するなど、広い応用が可能である。

【００３４】以上の計算式を利用すると、例えば図３に示す例のようにして関連キーワード自動抽出を行うことができる。この図３は関連キーワード自動抽出動作の処理手順の流れを示す図である。図３において、文書番号リスト３１が入力された単語統計情報管理部１７は、該当する文書番号（例えば００１０、０３４１等）に出現する単語およびその頻度を文書ごとに出力し、文書内単語統計情報３３，３４，３５を得る。同時に、ここで求められたすべての単語に対して、全文書中での統計情報３２を得る。次にこれらの統計情報３２、３３、３４、３５が単語ランキング部１８に渡される。単語ランキング部１８では、各種統計情報３２～３５を基に、例えば前記の式を利用して各単語の重要度を計算する。図３の場合だと、以下のようになる（ただし、Ｃを１、Ｎを１００００とする）。

ＩＤＦ（アプレット）＝　１－ｌｏｇ（８６／１００００）
＝　５．７５６
Ｓ（アプレット）　＝　２＊５．７５６＋６＊５．７５６）＊２
＝　９２．０９６
ＩＤＦ（インターネット）＝　１－ｌｏｇ（１１２９／１００００）
＝　３．１８１
Ｓ（インターネット）＝　（３＊３．１８１＋１＊３．１８１＋２＊３

．１８１）＊３
＝　５７．２５８
ＩＤＦ（ＣＧＩ）　＝　１－ｌｏｇ（７９／１００００）
＝　５．８４０
Ｓ（ＣＧＩ）　＝（４＊５．７５６）＊１
＝　２３．０２４
ＩＤＦ（ＷＷＷ）　＝　１－ｌｏｇ（６１５／１００００）
＝　３．７８９
Ｓ（ＷＷＷ）　＝（５＊３．７８９）＊１
＝　１８．９４５
ＩＤＦ（ＪＡＶＡ）　＝　１－ｌｏｇ（１６１／１００００）
＝　５．１２９　　６
Ｓ（ＪＡＶＡ）　＝（６＊５．１２９＋３＊５．１２９＋３＊５．１２９）＊３
＝　１８４．６４４
ＩＤＦ（ＳＵＮ）　＝　１－ｌｏｇ（３５／１００００）
＝　６．６５５
Ｓ（ＳＵＮ）　＝（６＊６．６５５）＊１
＝　３９．９３０
ＩＤＦ（スクリプト）＝　１－ｌｏｇ（８１３／１００００）
＝　３．５１０
Ｓ（スクリプト）　＝（５＊３．５１０）＊１
＝　１７．５５０

【００３５】単語ランキング部１８では以上のように求められた重要度により単語を整列し、整列後の単語リスト３７を得る。ここで、ランキングされた単語の上位３語を抽出するという指定になっているとすれば、単語リスト３７における上位３語である「ＪＡＶＡ」「アプレット」「インターネット」が関連キーワードとして抽出される。

【００３６】以上では辞書に登録された一単語を抽出の対象としてきたが、一般に単語だけでなく、単語の組でもよい。単語の組とは、名詞の連続により構成される複合語や、助詞「の」で結ばれる名詞の組、助詞「を」「が」で結ばれる名詞と動詞の組などを指す。これらの統計情報が単語と同様に事前に抽出できているのであれば、上記で示した手法がそのまま適用でき、単語の組を関連キーワードとして抽出することができる。

【００３７】なお、関連キーワード入力装置１６は、文書集合選定部１９および条件入力部２０を別構成としてもよい。特に文書集合選定部１９が検索式による文書検索手段を有する場合には、後出の図７に示すような別構成とすることで、文書検索装置による文書番号を入力として受け、出力される関連キーワードを文書検索装置の検索式入力部に反映させることができる。

【００３８】このように、本実施の形態によれば、対象となる文書のうちの一部である文書の部分集合が特定された際、当該部分集合に含まれる各文書に出現する各単語それぞれについて重要度を計算して重要度の順に整列し、整列された単語群のうちの一部を抽出して関連キーワードとすることで、動的かつ高速に対象となる文書の特性に即した関連キーワードを求めることができるという効果を持つ。

【００３９】また、上記のようにして得られた関連キーワードは、同一文書を対象とする文書検索装置への入力として利用することができ、その場合、対象文書の特性にあった的確なキーワードを再利用できるだけでなく、当該関連キーワードは必ず対象文書に含まれることが保証されるため、これを利用して検索した場合に必ず検索結果が得られるという効果も持つ。

【００４０】また、得られた関連キーワードを同一の対象文書集合または別の対象文書集合を対象とする文書検索装置への入力として利用することができ、その場合には、関連キーワード抽出の対象となった文書集合において特徴的であるキーワードをもとに、同一または別の文書集合を検索することができ、特に別の文書集合を検索対象とする文書検索装置の場合に、当該キーワードを異なった特性を持つ文書集合に対しても適用することができるという効果をもつ。

【００４１】また、抽出されたキーワードをユーザに提示して選択させるという構成とすることで、ユーザが再検索を実行する際、キーボードから再度検索条件を入力する代わりに、関連キーワードを、例えばマウスのクリックなど単純な操作で選択することが可能となり、再検索における操作を軽減して検索の効率を高めると同時に、検索の操作に不慣れなユーザでも簡単に利用できるという効果を持つ。

【００４２】また、抽出された関連キーワードにその重要度も付加して提示することにより、例えば検索条件との適合度を計算して文書をランキングするような文書検索装置において、検索条件中の各単語に対して重みを付与することができる文書検索装置であれば、抽出されたキーワードとその重要度をそのまま入力とすることで、より高精度の検索結果を得ることができるという効果を持つ。

【００４３】また、抽出された関連キーワードとその重要度の組を、例えばユーザの履歴として蓄積していくことにより、ユーザの興味の範囲や嗜好などをキーワードとその重みのベクトルとして表現することが可能となり、このベクトルを他の文書集合の検索に利用するなど、広い応用が可能であるという効果も持つ。

【００４４】（実施の形態２）次に、本発明の第２の実施の形態について実施の形態１に示したブロック図と同じ図１を利用して説明する。この第２の実施の形態では、文書集合選定部１９が２種類の文書集合Ａおよび文書集合Ｂを特定する。ここで、文書集合Ｂは文書集合Ａの部分集合となっている。例えば、ある検索式で検索を行った結果得られる文書集合Ａと、そのうちで関連する文書群としてユーザが指定した文書集合Ｂとが特定される場合や、文書の属性により特定された文書集合Ａと、その中でさらに検索式により絞り込まれた文書集合Ｂとが特定される場合などである。

【００４５】この場合、例えば以下の式により算出される単語の分布指標を当該単語の重要度に乗算するなどの手法により、単語の重要度を算出する。

ＤＩ（Ａ，Ｂ，Ｗ）＝　｛（ＮＡ／ＤＡ（Ｗ））＊（ＤＢ（Ｗ）／ＮＢ）｝

ただし、

ＤＡ（Ｗ）：部分集合Ａにおける単語Ｗの出現する文書数

ＤＢ（Ｗ）：部分集合Ｂにおける単語Ｗの出現する文書数

ＮＡ：部分集合Ａの総文書数

ＮＢ：部分集合Ｂの総文書数

【００４６】これは、部分集合Ｂにおいて高い頻度で出現し、かつ部分集合Ａにおける出現頻度が低いものほど高い値となる。上式において高い値となる語は部分集合Ａにおいて部分集合Ｂの弁別性に大きく寄与するものであり、部分集合Ｂをより特徴づけるキーワードであるといえる。例えば、図３に示す例において、文書番号リスト３１が部分集合Ｂであるとし、これを含む部分集合Ａ（総文書数１００とする）も同時に指定されている場合で、部分集合Ａ中の各単語の出現文書数が以下の通りであるとする。

ＤＡ（アプレット）　＝　１０
ＤＡ（インターネット）＝　２８
ＤＡ（ＣＧＩ）　＝　９
ＤＡ（ＷＷＷ）　＝　１４
ＤＡ（ＪＡＶＡ）　＝　２０
ＤＡ（ＳＵＮ）　＝　５
ＤＡ（スクリプト）　＝　１０

【００４７】この場合各単語の重要度Ｓ２（Ｗ）は、実施の形態１で説明した各単語の重要度Ｓ（Ｗ）に各単語の重みＤＩ（Ａ，Ｂ，Ｗ）を乗算した値となり、以下のように計算される。

Ｓ２（アプレット）　＝ ９２．０９６ ＊ ｛（１００／１０）＊（２／３）｝
＝ ６１３．９７３

Ｓ２（インターネット）＝ ５７．２５８ ＊ ｛（１００／２８）＊（３／３）｝
＝ ２０４．４９３

Ｓ２（ＣＧＩ）　＝ ２３．０２４ ＊ ｛（１００／９）＊（１／３）｝
＝ ８５．２７４

Ｓ２（ＷＷＷ）　＝ １８．９４５ ＊ ｛（１００／１４）＊（１／３）｝
＝ ４５．１０７

Ｓ２（ＪＡＶＡ）　＝ １８４．６４４ ＊ ｛（１００／２０）＊（３／３）｝
＝ ９２３．２２０

Ｓ２（ＳＵＮ）　＝ ３９．９３０ ＊ ｛（１００／５）＊（１／３）｝
＝ ２６６．２００

Ｓ２（スクリプト）　＝ １７．５５０ ＊ ｛（１００／１０）＊（１／３）｝
＝ ５８．５００

となり、重要度の順に整列すると

Ｓ２（ＪＡＶＡ）　＝　９２３．２２０
Ｓ２（アプレット）　＝　６１３．９７３
Ｓ２（ＳＵＮ）　＝　２６６．２００
Ｓ２（インターネット）＝　２０４．４９３
Ｓ２（ＣＧＩ）　＝　８５．２７４
Ｓ２（スクリプト）　＝　５８．５００
Ｓ２（ＷＷＷ）　＝　４５．１０７

の順となる。したがって、上位３語を関連キーワードとして抽出するのであれば、「ＪＡＶＡ」「アプレット」「ＳＵＮ」が関連キーワードとなる。

【００４８】上記の計算式は一例であり、部分集合Ｂにおいて高い頻度で出現し、かつ部分集合Ａにおける出現頻度が低いものほど高い値となるような他の計算式を利用してもよい。

【００４９】このように、本実施の形態によれば、特定された２種類の部分集合間における頻度分布の差異を考慮することにより、より高精度な関連キーワードを得ることができるという効果を持つ。

【００５０】（実施の形態３）次に、本発明の第３の実施の形態について実施の形態１に示したブロック図と同じ図１を利用して説明する。この第３の実施の形態では、文書集合選定部１９に各文書の重みを付与する機能を設ける。例えば、ユーザが文書を指定する場合に、各文書に対して関連度を指標として５段階の評価値を与える場合や、検索式による検索の結果得られる文書が検索式との適合度によりランキングされている場合に１位に１０点、２位に９点、といった重みを与える場合などである。単語ランキング部は各文書に付与された重みを、当該文書に含まれる単語に対して、例えば乗算するなどして加味し重要度算出を行う。なお、各文書に与える重みは負の値であってもよい。例えば、ユーザが文書を特定する際、関連する文書には２点、まったく関連しない文書には－１点を与える、という重み付与も許す。これにより、関連する文書にも関連しない文書にも含まれる（かつあまり一般的でない）語の重要度を低くすることができる。

【００５１】このように、本実施の形態によれば、特定した文書集合に含まれる各文書に対して重みを与えることにより、より重要な文書に含まれる単語ほど高い重要度となるような計算式とすることで、文書それぞれの重要度を勘案した高精度な関連キーワードが得られるという効果を持つ。

【００５２】（実施の形態４）次に、本発明の第４の実施の形態について説明する。図４は本発明の第４の実施の形態に係る関連キーワード自動抽出装置のブロック図である。この第４の実施の形態では、第１の実施の形態の構成に加えて閾値設定部２２を有して成り、この閾値設定部は単語統計情報管理部１７との間でデータの送受ができるようになっている。また、この実施の形態においては、単語統計情報管理部１７には閾値による単語除外機能が付与されている。かかる構成において、単語統計情報管理部１７は各単語の統計情報を出力する際、あらかじめ定められた閾値設定２２を参照し、極端に高頻度または低頻度の単語はその場で候補から除外して単語ランキング部１８に当該単語の情報を出力しない構成とすることができる。例えば、閾値１を「全文書の５０％以上に出現する単語」と設定し、閾値２を「１文書にしか出現しない単語」と設定することで、これらの単語が重要度算出に与える悪影響を事前に防ぐことができ、かつ処理の高速化を図ることができる。

【００５３】なおその際、単語の長さなど当該単語の特徴量に応じて、閾値を何種類かに設定してもよい。例えば、日本語の場合で「二文字以上の語は全体の５０％以上、一文字の語は全体の３０％以上」といった閾値設定を行うことで、各語の特性にあわせて除外する単語の範囲を設定する。

【００５４】このように、本実施の形態によれば、対象文書集合全体において出現度合いが高頻度または低頻度である単語をあらかじめ定められた閾値を考慮して除外することにより、キーワード抽出処理を高速化でき、かつ再利用の際に有効性の高い単語のみが選別できるという効果を持つ。

【００５５】（実施の形態５）次に、本発明の第５の実施の形態について説明する。図５は本発明の第５の実施の形態に係る関連キーワード自動抽出装置の構成を示すブロック図である。この第５の実施の形態に係る関連キーワード自動抽出装置は、第１の実施の形態において説明したような、文書全体の単語統計情報１４および文書内単語統計情報１５を管理する単語統計情報管理部１７、単語ランキング部１８、対象文書の部分集合を特定する文書集合選定部１９、および文書集合選定部１９への選定条件入力手段である条件入力部２０を有する基本構成に加えて、単語ランキング部１８と連動して単語の属性などの各種情報を利用することにより、抽出される関連キーワード群の質を向上させることを目的とするものである。図５において、符号２５は出現情報管理部、２６は単語属性情報管理部、２７は文字列包含関係判定部であり、これらの機能部は関連キーワード自動抽出装置２９に含まれて単語ランキング部と連動する。また２８は代表語付与部であり、この代表語付与部２８は単語ランキング部１８からデータを受けて関連キーワードを

出力する。また、関連キーワード自動抽出装置２９に対して、外部機能部として、対象文書集合１１からのデータを基に単語が出現する位置の情報を抽出する単語出現位置情報抽出部２３が設けられ、この単語出現位置情報抽出部２３からは出現位置情報２４が出力される。この出現情報は出現情報管理部２５へ送付される。

【００５６】かかる構成を有する本発明の第５の実施の形態について、その動作を説明する。この実施の形態の動作においては、まず対象となる文書集合１１に対し、辞書１２を利用して、前処理として動作する統計情報抽出部１３により、対象文書集合１１全体における単語の出現頻度・分布などの単語統計情報１４、および各文書ごとの当該文書に含まれる単語の統計情報である文書内単語統計情報１５を抽出しておく。同時に、必要があれば単語位置情報抽出部２３により、単語の出現位置情報２４も抽出しておく。図６は単語出現位置情報抽出部２３によって抽出された出現位置情報２４のデータ構造の一例を表すテーブル構成図である。出現位置情報は例えば図６に示すようなテーブルとして格納される。各文書ごとにその文書に出現する単語と出現位置（例えば文書の先頭からのバイトオフセット）、出現区分などが格納される。

【００５７】そして関連キーワード自動抽出動作に際しては、各単語に対して出現情報管理部２５に問い合わせを行い、当該単語の出現位置や出現文脈などの情報を得、これを重要度算出に加味する。例えば、検索対象とする文書すべてが、タイトル（または見出し）、サブタイトル、本文、といった要素から構成されている文書である場合、当該単語がこれら要素のうちいずれに含まれているかによって、

タイトルに含まれる場合には３点
サブタイトルに含まれる場合には２点
本文に含まれる場合には１点

といったような「重み」を各単語の重要度に乗算する、といった手法で重要度を算出する。

【００５８】あるいは、出現位置の情報を利用してもよい。例えば部分集合が検索式により特定される場合で、この検索式に含まれる単語が参照可能である場合、検索式に含まれる単語と、現在重要度計算の対象となっている単語との間の文字数が、

２文字以内なら３点
１０文字以内なら２点
１０文字以上なら１点

といったような「重み」を当該単語の重要度に乗算する、といった手法で重要度を算出することも可能である。

【００５９】また、本実施の形態の別の態様として、各単語に対して、単語属性情報管理部２６に問い合わせを行い、当該単語の品詞や分類など、その単語の属性を得、これを重要度算出に加味する。例えば、当該単語の品詞に着目し、

固有名詞ならば５点
普通名詞ならば４点
形容詞、形容動詞ならば２点
動詞、副詞ならば１点
その他自立語でないもの（助詞、助動詞など）ならば０点

といったような「重み」を各単語の重要度に乗算する、といった手法で重要度を算出することも可能である。

【００６０】また、本実施の形態の別の態様として、ある２つの単語間の文字列としての包含関係を判定する文字列包含関係判定部２７を用いて、抽出された単語同士、もしくはあらかじめ指定された単語群のうちの一単語と抽出された単語との間に包含関係があるか否かを判定し、包含関係があると判定された場合に、抽出する単語を制限する。ここであらかじめ指定された単語群とは、例えば部分集合の特定に検索式を利用した場合の検索式に含まれる単語などである。包含関係の判定においては、あらかじめ定められた設定により、以下の判定基準のいずれか一つ（または一つ以上）を満たす場合を包含関係と認定することができる。

（１）単語Ａと単語Ｂとが前方において一致しかつ単語Ａが単語Ｂより短い場合、（２）単語Ａと単語Ｂとが後方において一致しかつ単語Ａが単語Ｂより短い場合、（３）単語Ａが単語Ｂの部分でありかつ前方、後方ともに一致しない場合、（４）単語Ａと単語Ｂとの関係が（１）～（３）のいずれかを満たし、かつ単語Ｂの構成要素と完全に一致する場合、

【００６１】例えば、（１）の基準では「東京都」に対する「東京」が部分語と判定される。以下、同様にして、（２）の基準では「新発売」に対する「発売」が、（３）の基準では「大感謝祭」に対する「感謝」が、それぞれ部分語と判定される。（４）の基準は、英語における部分語判定の際に重要であり、この基準に従えば "artificial intelligence" に対して "art" や "tel l" は部分語とはならないが、"artificail" や "intel ligence" は部分語と判定される。

【００６２】上記基準により、部分語関係にあると判定された２つの語について、そのどちらを関連キーワードとして採用するかについても、以下のいずれかの基準（あらかじめ設定されているものとする）に従う。

（１）　長単位の単語を採用する
（２）　短単位の単語を採用する
（３）　重要度の高い単語を採用する
（４）　短単位の単語および長単位の単語と短単位の単語との差分を採用する

【００６３】例えば、単語「東京都」が重要度１０で、単語「東京」が重要度７でそれぞれ抽出され、かつ両者に部分語関係が成立した場合、（１）の基準に従うと文字列として長い「東京都」が採用され、（２）の基準に

従うと文字列として短い「東京」が採用され、（３）の基準に従うとより重要度の高い「東京都」が採用されることになる。（４）の基準は、例えば単語 "artificial intelligence" と "artificial" との間に部分語関係が成立した場合に、"artificial" および "intelligence" を関連キーワードとして採用するものであり、主に英語文書において効果的である。

【００６４】あらかじめ指定された単語群との間に部分語関係が成立する単語の場合、（３）以外の手法が利用できる。その場合、「短単位（もしくは長単位）であれば関連キーワードとして採用しない」といった処理となる。抽出された単語同士に部分語関係が成立する場合には、いずれの手法も利用可能である。

【００６５】また、本実施の形態の別の態様として、抽出された関連キーワード群を、各語の属性や統計情報を利用して分類して提示する。語の属性として品詞を利用すると、例えば固有名詞とそれ以外に分類して提示することができる。あるいは、語の属性としてシソーラス辞書を利用し、各語をシソーラスにおける分類に対応する形で分類して提示することも可能である。また、統計情報を利用した分類とは、例えば特定された文書集合における各語の出現文書数により分類する手法などがあげられる。その場合、例えば「出現文書数が文書集合の８割以上であるか否か」といった基準で分類することで、その語が再検索に利用される際の絞り込みの効果を事前に確認することができる。なお、分類にあたり語の属性としてシソーラス辞書を利用する場合、分類された単語群に対して、シソーラスの上位ノードに相当する語を代表語として与え、単語群をその語で代表させることも可能である。同様に、単語の統計情報１４を利用する場合には、分類された単語群において、例えば最も出現頻度の高い語を代表語として採用してもよい。

【００６６】このように、本実施の形態によれば、単語が出現した位置の情報を利用することで、文書の構造や単語間の距離の情報を考慮した関連キーワードの抽出が行なえ、高精度な関連キーワード抽出が可能となるという効果を持つ。

【００６７】また、単語の品詞など、各単語の属性情報を考慮することにより、各属性の特徴に応じた関連キーワードの抽出が行なえ、高精度な関連キーワード抽出が可能となるという効果を持つ。

【００６８】また、単語間の文字列としての包含関係を考慮することにより、同じような意味や用途である単語を排除して関連キーワードの抽出が行なえ、関連キーワード全体としての冗長性を抑えることができるという効果を持つ。

【００６９】また、抽出された関連キーワードを分類し、必要があれば各分類に対応する代表語を設定することで、抽出されたキーワードの一覧性や傾向、再利用の際の有効性などをあらかじめ確認して関連キーワードの抽出が行なえ、関連キーワードとしての使いやすさを向上することができるという効果を持つ。

【００７０】（実施の形態６）次に、本発明の第６の実施の形態について説明する。図７は本発明の第６の実施の形態に係る文書検索装置の構成およびこれと関連キーワード自動抽出装置とを組み合わせて実現した文書検索システムの構成を示すブロック図である。この文書検索装置４１は、前記第１，第２，第３、第４または第５の実施の形態に係る関連キーワード自動抽出装置と連携して動作するものである。

【００７１】本実施形態における文書検索装置４１は、文書検索に必要な条件式を入力する検索条件入力部４４と、入力された検索条件にしたがって文書の検索を行なう文書検索部４５と、文書検索部４５において検索された文書について入力された検索式と文書との間の適合度を計算する文書ランキング部４６とを有して成る。この文書検索装置４１は、連携して動作する関連キーワード自動抽出装置４８と同一の対象文書集合１１を検索対象とするものであり、単語統計情報抽出に利用するのと同一の辞書１２を利用して、あらかじめ索引生成部４２により作成された文書検索用の索引４３を利用して検索を行う。また、本実施形態における関連キーワード自動抽出装置４８は、文書集合選定部４７を別構成としたものであり、関連キーワード自動抽出装置４８には、文書集合選定部４７が特定した文書の部分集合の各要素に対応する文書の識別子の集合（一意である文書番号のリストなど）が入力される。

【００７２】以上の構成を備えた本実施の形態について、その動作を説明する。最初に検索条件入力部４４に入力された検索条件をもとに、文書検索部４５が検索用索引４３を参照して検索条件に適合する文書を特定する。ここで得られた文書集合をそのまま検索結果文書５０としてもよいが、さらに文書ランキング部４６において、入力された検索式と文書との間の適合度を計算して適合度の高い順に文書を整列したものを検索結果とする、といった構成にしてもよい。こうして得られた検索結果の文書集合５０は、ユーザに検索結果として返すのと同時に、文書集合選定部４７に渡される。文書選定部４７では、文書ランキング部４６から渡された文書集合のすべてまたは一部を関連キーワード自動抽出装置４８への入力として採用する。文書が適合度の順にランキングされているのであれば、検索結果の文書集合のうち例えば上位１０文書を選定する、という構成にしてもよい。また、あらかじめ文書ごとに付与された属性情報が利用できるのであれば、これを利用して特定の属性値を持つ文書のみを選定する、という構成としてもよい。

【００７３】文書集合選定部４７により特定された文書の部分集合は関連キーワード自動抽出装置４８に送られ、前記第１，第２，第３，第４または第５の実施の形態に示したような手順で関連キーワード群４９を抽出す

る。こうして得られた関連キーワード群４９は検索条件入力部４４に戻され、ユーザに提示される。ユーザは提示された関連キーワード群から必要なものを選択して新たな検索条件とし、再度検索を実行させることができる。これにより、本実施の形態によれば、関連キーワード自動抽出装置によって上記のようにして得られた関連キーワードは、同一文書を対象とする文書検索装置への入力として利用することができ、その場合、対象文書の特性にあった的確なキーワードを再利用できるだけでなく、当該関連キーワードは必ず対象文書に含まれることが保証されるため、これを利用して検索した場合に必ず検索結果が得られるという効果も持つ。

【００７４】（実施の形態７）次に、本発明の第７の実施の形態について説明する。図８は本発明の第７の実施の形態に係る文書検索装置の構成およびこれと関連キーワード自動抽出装置とを組み合わせて実現した文書検索システムの構成を示すブロック図である。この文書検索装置５１は、第６の実施の形態に係る文書検索装置４１と同様、前記第１，第２，第３、第４または第５の実施の形態に係る関連キーワード自動抽出装置と連携して動作するものである。

【００７５】本実施形態における文書検索装置５１は、文書検索に必要な条件式を入力する検索条件入力部５４と、入力された検索条件にしたがって文書の検索を行なう文書検索部５５とを有して成る。本実施形態における文書検索装置５１は、連携して動作する関連キーワード自動抽出装置５２とは異なる対象文書集合５６を検索対象とするものであり、文書検索部５５が対象文書集合５６に接続される構成となっている。なお、その検索手法についての詳細は問わない。

【００７６】以上の構成を備えた本実施形態における動作について、以下説明する。最初に指定された条件にしたがって関連キーワード自動抽出装置５２が動作し、関連キーワード群５３を出力する。文書検索装置５１における検索条件入力部５４は、関連キーワード群５３を入力としてユーザに提示し、ユーザは提示された関連キーワードのうち必要なもののみを選択して、検索対象となる対象文書集合５６に対する検索を実行し、検索結果文書５７を得ることができる。

【００７７】このように、本実施の形態によれば、関連キーワード自動抽出装置５２によって得られた関連キーワードを同一の対象文書集合または別の対象文書集合を対象とする文書検索装置５１への入力として利用することができ、その場合には、関連キーワード抽出の対象となった文書集合において特徴的であるキーワードをもとに、同一または別の文書集合を検索することができ、特に別の文書集合を検索対象とする文書検索装置の場合に、当該キーワードを異なった特性を持つ文書集合に対しても適用することができるという効果をもつ。

【００７８】【発明の効果】以上説明したように、本発明によれば、関連キーワード自動抽出装置を、文書の部分集合を特定する文書集合選定部と、対象文書全体または個々の文書ごとに出現する単語とその統計情報を管理する単語統計情報管理部と、文書の部分集合に出現する各単語の重要度を算出して重要度の順に整列する単語ランキング部ととにより構成したため、文書全体、および特定された文書部分集合における各単語の統計情報を高速に求めることが可能であり、特定された文書集合に出現する各単語を、その重要度に基づいて高速にランキングし、その一部を関連キーワードとして提示することができる。

【００７９】また、前記構成に加えて、単語の属性情報や文書内の出現位置を管理する手段などを設けることにより、単語の重みを変化させ、あるいはランキング後の単語群から特定の条件を満たす単語を削除することで、抽出される単語群の関連語としての精度を向上させることができる。また、抽出された単語群を、語の属性や統計的性質により分類することで、よりわかりやすい関連キーワード提示を行なうことができる。

【００８０】さらに、関連キーワード自動抽出装置と連携した文書検索装置を含む文書検索システムを構成し、抽出された関連キーワードを入力として再利用することにより、抽出された関連キーワードが対象文書の特性に合ったものであり、かつ検索対象が同一の文書群であるならば、そのキーワードによって検索結果が少なくとも１件以上得られることが保障されるため、効率的かつ容易に再検索を行なうことができる等の効果が得られる。

【図面の簡単な説明】

【図１】本発明の第１乃至第３の実施の形態に係る関連キーワード自動抽出装置の構成を示すブロック図

【図２】（ａ）　前記実施の形態における単語統計情報の構造を示すテーブル構成図

（ｂ）　前記実施の形態における文書内単語統計情報の構造を示すテーブル構成図

【図３】前記実施の形態における関連キーワード自動抽出動作の処理手順の流れを示す図

【図４】本発明の第４の実施の形態に係る関連キーワード自動抽出装置の構成を示すブロック図

【図５】本発明の第５の実施の形態に係る関連キーワード自動抽出装置の構成を示すブロック図

【図６】前記実施の形態における単語出現位置情報抽出部によって抽出された出現位置情報のデータ構造の一例を表すテーブル構成図

【図７】本発明の第６の実施の形態に係る文書検索装置の構成構成およびこれと関連キーワード自動抽出装置とを組み合わせて実現した文書検索システムの構成を示すブロック図

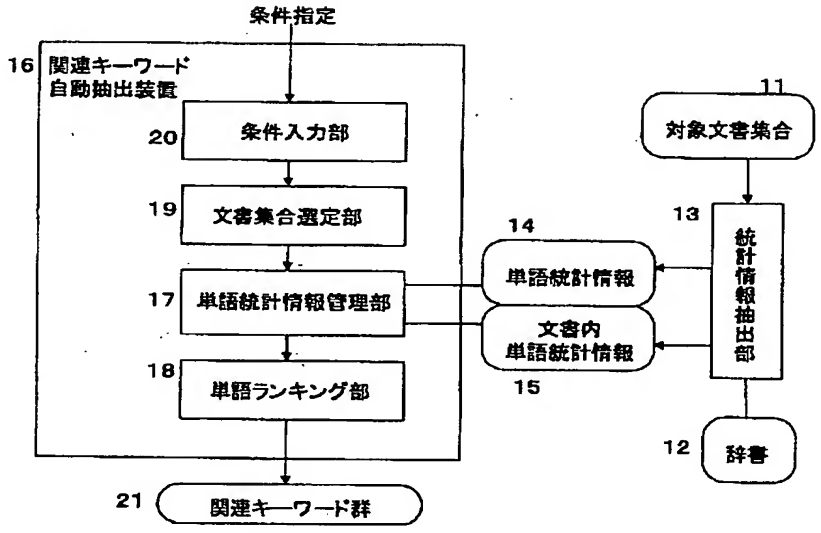
【図８】本発明の第７の実施の形態に係る文書検索装置の構成およびこれと関連キーワード自動抽出装置とを組み合わせて実現した文書検索システムの構成を示すブロ

ック図

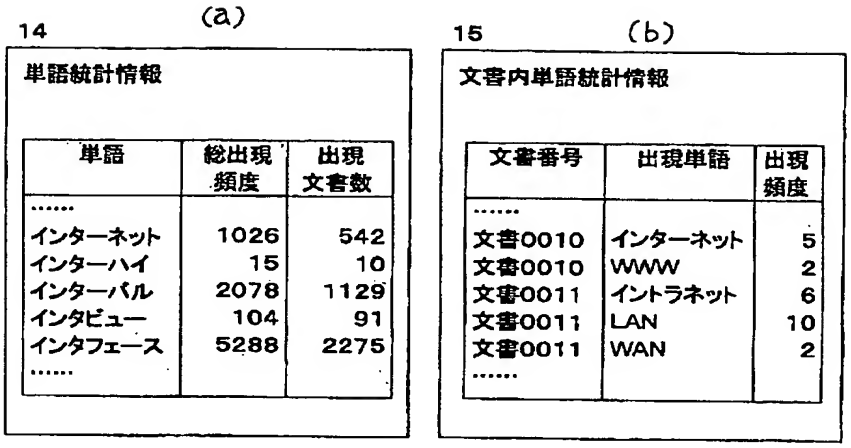
【符号の説明】

１１、５６　対象文書集合
１２　辞書
１３　統計情報抽出部
１４　単語統計情報
１５　文書内単語統計情報
１６、２９、４８、５２　関連キーワード自動抽出装置
１７　単語統計情報管理部
１８　単語ランキング部
１９　文書集合選定部
２０　条件入力部
２１、４９、５３　関連キーワード群
２２　閾値設定
２３　単語出現位置情報抽出部
２４　出現位置情報
２５　出現情報管理部
２６　単語属性情報管理部
２７　文字列包含関係判定部
２８　代表語付与部
４１、５１　文書検索装置
４２　索引生成部
４３　検索用索引
４４、５４　検索条件入力部
４５、５５　文書検索部
４６　文書ランキング部
４７　文書集合選定部
５０、５７　検索結果文書

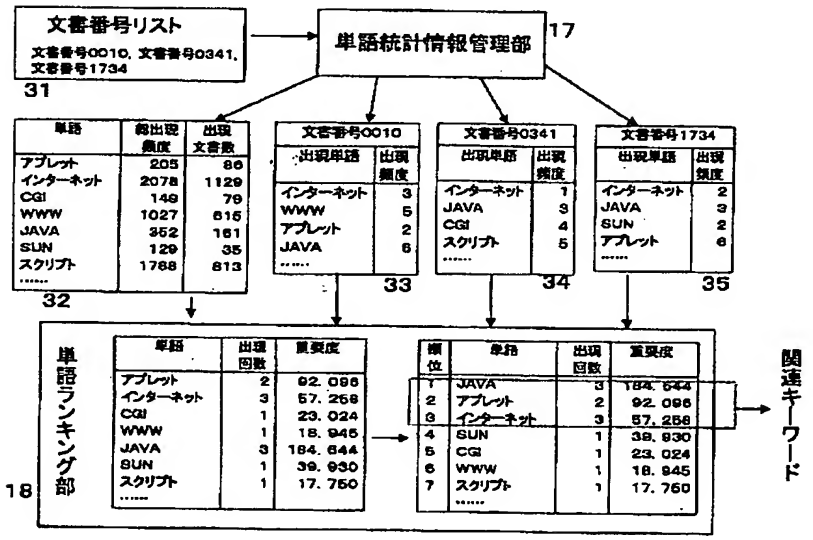
【図１】



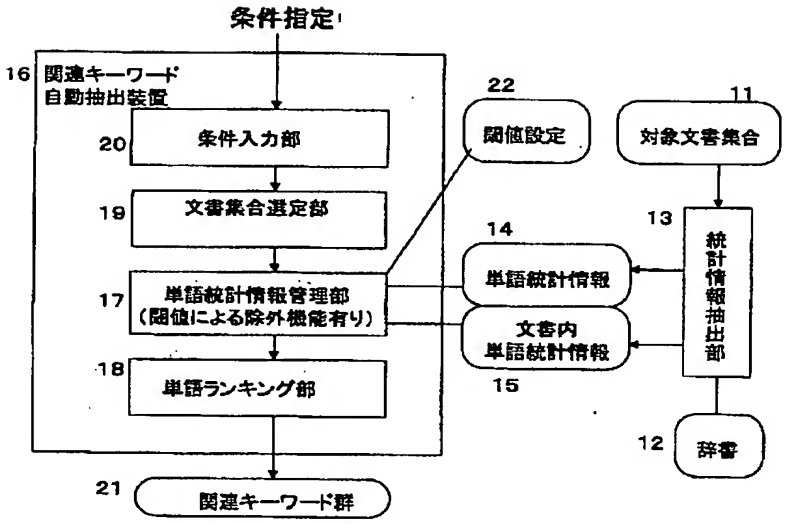
【図２】


(a)
14
単語統計情報
単語
総出現頻度
出現文書数
インターネット 1026 542
インターハイ 15 10
インターバル 2078 1129
インタビュー 104 91
インタフェース 5288 2275
(b)
15
文書内単語統計情報
文書番号
出現単語
出現頻度
文書0010 インターネット 5
文書0010 WWW 2
文書0011 イントラネット 6
文書0011 LAN 10
文書0011 WAN 2

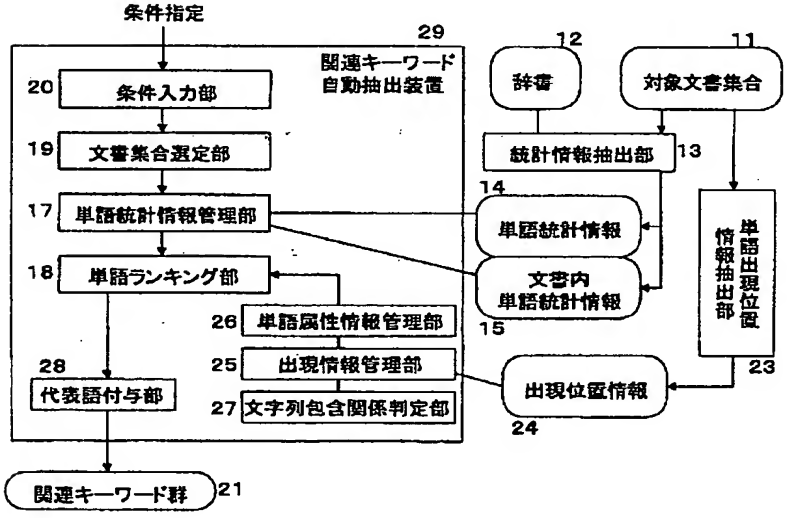
【図３】


文書番号リスト
文書番号0010, 文書番号0341, 文書番号1734
31
単語統計情報管理部
17
32
33
34
35
単語ランキング部
18
関連キーワード

【図４】


条件指定
16 関連キーワード自動抽出装置
20 条件入力部
19 文書集合選定部
17 単語統計情報管理部（閾値による除外機能有り）
18 単語ランキング部
21 関連キーワード群
22 閾値設定
11 対象文書集合
13 統計情報抽出部
14 単語統計情報
文書内単語統計情報
15
12 辞書

【図５】


条件指定
29 関連キーワード自動抽出装置
20 条件入力部
19 文書集合選定部
17 単語統計情報管理部
18 単語ランキング部
26 単語属性情報管理部
25 出現情報管理部
27 文字列包含関係判定部
28 代表語付与部
関連キーワード群 21
12 辞書
11 対象文書集合
統計情報抽出部 13
14 単語統計情報
文書内単語統計情報
15
単語出現位置情報抽出部 23
出現位置情報 24

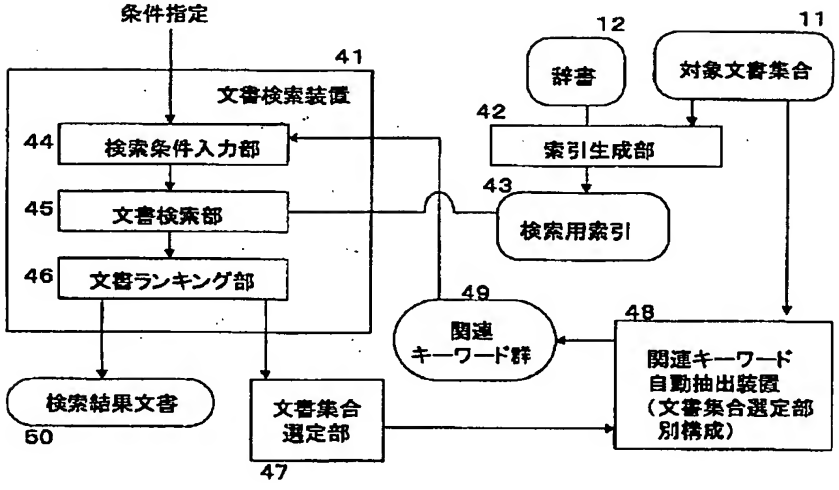
【図６】

24

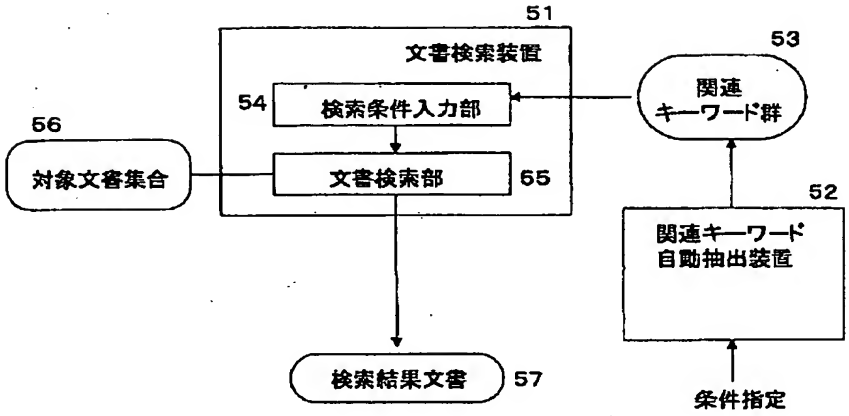
単語出現位置情報

文書番号	出現単語	出現位置	出現区分
……			
文書0010	インターネット	10368	タイトル
文書0010	WWW	10384	タイトル
文書0010	近年	10390	本文
文書0010	インターネット	10396	本文
文書0010	成長	10412	本文
……			

【図７】



【図８】



フロントページの続き

(72)発明者　野　本　昌　子
大阪府門真市大字門真1006番地　松下電器産業株式会社内

(72)発明者　稲　葉　光　昭
大阪府門真市大字門真1006番地　松下電器産業株式会社内

(72)発明者　福　重　貴　雄
大阪府門真市大字門真1006番地　松下電器産業株式会社内